SHAPING SHANGRI-LA

*'Relict' landscapes of Tibet and Yunnan are not
time capsules — they are forming today* **PAGES 442 & 526**

9 770028 083095

Highway to health

Africa has an ambitious and welcome plan for a continent-wide centre for disease control — but if the agency is to live up to its promise, it will need substantially better resources.

In light of the ongoing Ebola epidemic in West Africa, it is entirely sensible that the African Union intends to set up a pan-continental organization to guard against the ravages of infectious disease. The African Centres for Disease Control and Prevention (ACDC), as the agency is called, will be modelled not on the massive US CDC, but on the smaller European Centre for Disease Prevention and Control (ECDC). Although the goal is excellent and the effort should be cautiously welcomed, the plans for the African agency are woefully inadequate. In terms of funds and staff — at least initially — it will be in no position to achieve its lofty ambitions.

There is no exact definition of a CDC other than a public-health agency, often with a focus on infectious diseases. The US CDC was created in 1946 with around 400 staff members and an annual budget of some US$10 million (worth $120 million today), largely to tackle malaria in the US south. It now has a staff of 15,000 and a whopping $7-billion annual budget, covering the entire spectrum of health issues — from detecting and responding to infectious-disease outbreaks worldwide, to control and prevention of non-communicable diseases, bioterrorism, workplace injuries and environmental health threats.

The past 15 years have seen a proliferation of CDCs and equivalent organizations worldwide, in part because of awareness of the need to improve responses to infectious-disease threats, prompted in particular by the 2003 epidemic of severe acute respiratory syndrome (SARS). But their organizational models are diverse. The Public Health Agency of Canada — created in 2004 after SARS outbreaks revealed weaknesses in the country's response — is a smaller version of the US CDC, with 2,400 employees and a budget of Can$633 million (US$517 million).

The ECDC is different. Unlike the US and Canadian organizations, the Stockholm-based agency has no in-house laboratories. It acts instead as a coordinator, drawing on a network of research labs and national health-protection bodies to reinforce disease surveillance and response. The highly respected agency has a relatively modest budget of €60 million (US$64 million), and 300 staff members.

The ACDC, to be based for now in Addis Ababa, would coordinate national resources and labs, as well as pan-African networks including the African Field Epidemiology Network and the African Network for Drugs and Diagnostics Innovation. It would focus on infectious diseases. But the similarities with the ECDC end there.

The ACDC's budget for July 2015 to December 2016 calls for just US$6.9 million, and the centre would initially have just 11 staff members, including management and 5 epidemiologists. These resources are hopelessly inadequate to tackle the agency's long list of stated ambitions, which include not only helping to prevent, detect and respond to disease outbreaks, but also strengthening health systems and conducting regional and national risk assessments. The African Union has declined to say what funding it hopes the ACDC will attract beyond this start-up phase. Its backers need to commit funding now for at least five to ten years — experts say that $60 million and 300 staff members would be the absolute minimum needed.

That Africa seems to want to assume political ownership of its response to public-health emergencies, rather than relying largely on outside agencies and support, can only be welcomed. Yet there is a risk that the creation of this skeletal agency might provide an excuse for complacency and inaction by politicians in Africa and globally. And the underlying problems that leave many countries vulnerable to disease outbreaks will not be solved by an African health agency alone, however robust and however desirable.

> *"That Africa wants to assume political ownership of its response to public-health emergencies can only be welcomed."*

As the Ebola epidemic has cruelly highlighted, what Africa lacks most are proper health systems and labs. There is a drastic shortage of health-care workers, with only a few dozen physicians per million people. Yet there are no serious national or international plans to reinforce public-health capacities.

An alphabet soup of organizations involved in global health and disease detection and prevention has developed over the past 20 years, yet proved incapable of reacting promptly when the Ebola threat emerged. These groups' often overlapping and duplicated efforts must be better coordinated to prevent and mitigate future health crises. But there is a danger that with concern over Ebola fading, governments will lapse back into old habits, and that the stark lessons of the epidemic will not be taken fully on board. ∎

Decoupled ideals

'Ecomodernist Manifesto' reframes sustainable development, but the goal remains the same.

Humanity faces two fundamental challenges this century. The first is to lift billions of people out of poverty and give them the opportunity to live healthy and dignified lives. The second is to ensure that this development does not destabilize the climatic and ecological systems that have enabled the rise of humans and other life on Earth. The problem is that these two goals are increasingly at odds.

Reconciling the twin imperatives of conservation and development is not easy. 'Sustainable development' is a catchphrase that neatly defines what the world must ultimately achieve, but nobody knows precisely what it looks like at full scale. Later this year, governments will finalize a set of sustainable development goals to guide international aid (see also page 432), and in December global leaders will gather to discuss the latest climate agreement at a summit in Paris. Any deal will be burdened

by inevitable compromises that allow space for polluting development as the world seeks better and cheaper solutions.

The latest attempt to create a framework for thinking about this dilemma comes from 18 environmental activists and academics, who published an 'Ecomodernist Manifesto' last week (see go.nature.com/f89sls). The essay paints a hopeful picture of technological progress while placing importance on the kind of intensive development that has characterized humanity's rise so far. Only by concentrating our impact within the urban, industrial and agricultural context can we achieve a "good Anthropocene", or age of human influence, the authors argue.

Coal, oil and natural gas have improved many lives, and the essay points out that the long arc of development is already tending towards better, cleaner and more-efficient energy technologies — just not fast enough. At least in the short term, the authors contend, poor countries are right to focus on improving the lives of their citizens, even if that means expanding their use of fossil fuels until cheaper and cleaner solutions are available. These ideas are framed in terms of a larger "decoupling of humanity from nature". What this means, precisely, is left to the imagination, but the essay also underscores the role of modern agriculture, which has freed up labour, enabled the rise of cities and reduced the amount of land that we need to feed humanity. Rather than lament this trend, the authors argue that it must be encouraged and hastened.

The essay stands in sharp contrast to the gloomy outlook often provided by environmentalists and scientists. A little scepticism is warranted. For the long haul, the authors place faith in a new generation of solar cells combined with efficient energy-storage technologies, advanced nuclear fission — and even fusion energy. In the medium term, hydropower could play a part, in the same way that technologies for the capture and sequestration of carbon could improve fossil fuels over this time scale. The authors focus on large-scale power generation, but may be too quick to write off current wind and solar technologies, which can have a useful role in reducing demand for centralized power today. The wise deployment of efficient bioenergy resources may also be helpful, in tandem with agricultural intensification.

> *"Governments cannot write people out of the equation."*

It is not yet clear what the climate fix will look like. What is clear is that governments need to invest in a portfolio of energy research, development and demonstration. They must implement strong climate policies that will push companies towards technologies that produce less air pollution and fewer greenhouse-gas emissions. They need to invest in agricultural research to secure the necessary food crops, and provide farmers and ranchers with the tools required to maximize production. And they need to set limits on the land that can be developed.

Governments cannot write people out of the equation, and hard choices will have to be made. But the first step is to point everybody in the right direction. Human ingenuity takes many forms, and we may yet surprise ourselves. ∎

More from Moore

Moore's law is approaching physical limits: truly novel physics will be needed to extend it.

Hail Gordon Moore: 19 April marked the famous prediction by the (less) famous man that the late twentieth century would herald massive increases in computing power, stimulating the technological age.

Electronics and information technology now touch almost every aspect of life. Kicking off with the invention of the integrated circuit in 1958, the continuing electronics revolution is, in large part, down to the technology industry's faithful compliance with what came to be known as Moore's law.

In 1965, Moore, a chemist turned electronic engineer, noticed that in the years since the first integrated circuits were built, engineers had managed to roughly double the number of components, such as transistors, on a chip every year. He also predicted that the rate of component shrinkage — which he later revised to a doubling every two years — would continue for at least another decade.

The semiconductor industry never looked back. It has continued to shrink transistors and produce computer chips that combine increasingly high performance and functionality.

For the first few decades, the semiconductor industry met Moore's law mainly through feats of engineering genius and gigantic strides in manufacturing processes. But the key role of fundamental science is also worth remembering, especially as researchers today seek ways to maintain the rate of progress.

The invention of the transistor at Bell Laboratories in Murray Hill, New Jersey, in the 1940s was firmly based on the development of semiconductor band theory. And scientific breakthroughs played an important part in the subsequent developments of technology. Notably, in 1970, the Russian physicist Nikolay Basov and collaborators developed excimer lasers that would later be used to etch tiny circuit patterns on the silicon wafers from which chips are made.

The 1990s called for further innovation. Until then, as transistors became smaller, their speed and energy efficiency increased. But when the components reached around 100 nanometres across, miniaturization began to have the opposite effect, worsening performance. Chip-makers such as Intel, which Moore co-founded, and IBM again looked to basic science to improve the performance of transistor materials. Major help came from condensed-matter physicists. They had known for decades that the ability of silicon to conduct electricity improves substantially when its crystal lattice is stretched — for instance, by layering it on another crystal in which the atoms have a different spacing. Engineers introduced strained silicon into chips in the 2000s, and Moore's law stayed true for several more years.

State-of-the-art microprocessors now have transistors that are just 14 nanometres wide, and Moore's law is finally approaching the ultimate physical limits. Waste heat in particular has become a source of concern. It has already caused one form of Moore's law — the exponential acceleration of computer 'clock speed' — to grind to a halt. Power-hungry chips also limit the ability of mobile devices to survive more than a few hours between charges.

The introduction of advanced materials such as hafnium oxide, which provides insulation even when it is just a few atomic layers thick, has managed to keep chips a bit cooler. Heroic efforts might yet bring one or two more generations of smaller transistors, down to a size of perhaps 5 nanometres. But further improvements in performance will require fundamentally new physics.

Where are we headed? Transistors that use quantum tunnelling, perhaps? Or those in which currents transport quantum spin rather than electric charge? Labs around the world are experimenting with approaches and materials that could drastically cut energy consumption. One avenue that could be exploited is the inherent stability of the collective 'topological' properties of atoms: a modern twist on the ancient practice of encoding information by tying knots. Some researchers are trying out radical 'neuromorphic' circuit architectures inspired by the plasticity of the brain's neuronal networks.

A principle that works well in a physics lab will not necessarily translate into something that can be mass-produced. And inevitably, most of today's attempts will lead nowhere. Society should have confidence, however, that somewhere, somehow, basic science will provide a way to maintain progress. Moore should be proud that we have not yet found the exception that proves his law. ∎

↻ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunqv

Why I teach evolution to Muslim students

Encouraging students to challenge ideas is crucial to fostering a generation of Muslim scientists who are free thinkers, says **Rana Dajani**.

Certain problematic attitudes towards science have been imported into Muslim societies as a part of rapid globalization and modernization — the rejection of the theory of evolution, for example. But this also offers an opportunity.

I teach evolution to university students in Jordan. Almost all of them are hostile to the idea at first. Their schoolteachers are likely to have ignored or glossed over it. Still, most students are willing to discuss evolution, and by the end of the course, the majority accept the idea. If Muslim students can challenge ideas on such a controversial academic topic, then they can also approach other aspects of their lives by questioning — and not just blindly accepting — the status quo. These tools and attitudes are crucial to the development of their personalities and to becoming responsible citizens.

Students in my classes often get a shock. I wear a hijab, so they know that I am a practising Muslim, yet they hear me endorsing evolution as a mechanism to explain diversity and the development of species, and citing Charles Darwin as a scientist who contributed to our understanding of the emergence and diversification of life on Earth. I am almost always the first Muslim they have met who says such things.

Some students complained to the university that I was preaching against Islam, but university officials were satisfied when I showed them that evolution featured in the university's approved textbooks and that what I teach in my lecture comes straight from these texts. I commended the students who complained for their courage in supporting what they believed, and offered to sit down and discuss their concerns with them.

In teaching, I offer a detailed explanation of the natural evolution of plants and artificial breeding. Later, we discuss antibiotic resistance, influenza vaccines and the development of HIV drugs. After these discussions, most students are willing to accept evolution as a mechanism for the emergence of all species except humans. Many quote evidence from the Koran that is interpreted to mean that Adam — and so humans — were created spontaneously. Human evolution remains taboo because the students are not ready to relinquish the concept that humans were created differently. I remind them that Muslims are warned against arrogance, and that humans are only part of creation.

Muslim scholars such as Hussein al-Jisr and Ahmad Medhat in the 1880s supported evolution. Before Darwin, al-Jahiz and others proposed rudimentary evolutionary theories in the ninth century. I point out that the apparent controversy over evolution and Islam arose only in the twentieth century, when Darwin's ideas became associated with colonialism, imperialism, the West, atheism, materialism and racism. Muslim religious scholars gradually took a stand against evolution, which the public adopted. The scholars used Christian creationist arguments to support their stance, transferring the Western war between science and religion to Islam.

Some of my students argue that to accept evolution means denying the existence of God. I say that evolution does not discuss the origins of the Universe. No one yet understands this beginning. To me, the beginning was God. After the beginning, the rules of logic and science led to the development of the Universe and beyond.

In my experience, many Muslims are happy to consider this. I have encountered several Muslim scientists who agree with my stance, but do not say so publicly because they fear being labelled as troublemakers. Some religious scholars also agree, but they wish to go about changing opinions gradually, so as not to raise defences and slow progress.

My take, as a Muslim scientist, is that the Koran asks humans to observe and contemplate the world while celebrating the pursuit of knowledge. It does not validate scientific findings. Science allows us to question and discover how the world works and the Koran provides the moral guidelines for doing so. If an apparent contradiction arises between a scientific finding and an interpretation of the Koran, then we can turn to both science itself (which is evolving) and the interpretation of the Koran (which is not impartial, because it is a human exercise) to account for the discrepancy. This is an ongoing and fluid process, and is part and parcel of the purpose of life for Muslims.

Whether or not a student accepts human evolution makes no difference to how I mark their exam paper. As educators, our objective is to help students to become independent thinkers. I do not want my students to write that they accept evolution just to pass an exam. I want them to show the argument they used to reach their conclusion, even if that conclusion rejects human evolution. Otherwise, I am doing what the people who decry evolution are doing: forcing an opinion on them.

My aim is to teach students to develop a rational methodology for assessing the natural world and to come up with their own opinions, hypotheses and theories and not to copy others. This becomes a call for new ways of thinking: a journey to pursue knowledge, which is one of the core tenets of Islam. If we succeed in that endeavour, we will contribute to the creation of a generation of Muslim scientists who are free thinkers. ■

> ## AS EDUCATORS, OUR OBJECTIVE IS TO HELP STUDENTS TO BECOME INDEPENDENT THINKERS.

Rana Dajani *is associate professor of molecular biology at Hashemite University in Zarqa, Jordan, and a visiting scholar at the Faraday Institute, University of Cambridge, UK.*
e-mail: rdajani@hu.edu.jo

A. AWAD

Sweet potato is already a GM crop

The sweet-potato genome contains genes from bacteria, so is an example of a naturally occurring genetically modified (GM) plant.

While combing through the genome of the domesticated sweet potato (*Ipomoea batatas*), Jan Kreuze of the International Potato Center in Lima, Peru, and his colleagues stumbled on genes that originated from one or more *Agrobacterium* species. These bacteria infect plants and transfer DNA into their hosts' genomes, so are used by researchers to genetically modify many agricultural crops. The two stretches of bacterial DNA found in the sweet potato are expressed in various tissues, and one of the sequences is found only in cultivated sweet potatoes — not in closely related wild strains.

These genes could encode traits that were desirable for domestication of the crop, and were probably transferred into an ancient sweet-potato ancestor, the authors say.
Proc. Natl Acad. Sci. USA
http://dx.doi.org/10.1073/pnas.1419685112 (2015)

Antarctic ice shelf nears its demise

The remains of an ice shelf that collapsed spectacularly in 2002 may be headed for total disintegration.

The break-up of the Larsen B ice shelf was one of the largest and fastest melting events ever seen by glaciologists. Ala Khazendar of the Jet Propulsion Laboratory in Pasadena, California, and his co-workers studied satellite and airborne radar measurements of the heights and flow speeds of glaciers feeding the remaining, southern section of Larsen B. They found that two of the glaciers thinned by 15–20 metres between 2002 and 2011, and one of them increased its flow speed by 55% from 1997 to 2012.

Modelling suggests that when most of Larsen B disintegrated, the remaining glaciers had less ice to buttress them in place, causing them to destabilize.
Earth Planet. Sci. Lett. 419, 199–210 (2015)

ROBERT CAPUTO/AURORA PHOTOS

Bacterial bonanza far from the West

Members of an isolated Amazon tribe in Venezuela (pictured) have the most diverse gut bacteria ever documented in humans.

Maria Gloria Dominguez-Bello at New York University School of Medicine analysed oral, faecal and skin bacteria from 34 Yanomami villagers who had never met anyone from the West until 2009. The researchers found almost double the genetic diversity in the faecal bacteria compared with the microbiomes of US residents. The faecal and oral bacteria also carried genes for antibiotic resistance, even though the people had no known exposure to the drugs.

In a separate study, Jens Walter at the University of Alberta in Edmonton, Andrew Greenhill at Federation University Australia in Churchill and their colleagues looked at the gut microbiomes of people from two rural areas of Papua New Guinea. They found that the bacteria were extremely diverse and included lineages not seen in US populations. Using ecological modelling, the authors attributed the increased diversity to low sanitation levels, which increased the swapping of bacteria between people.
Science Adv. 1, e1500183 (2015);
Cell Rep. http://doi.org/3vh (2015)

Some mutations in cancer arrive late

Subpopulations of tumour cells can harbour unique mutations that crop up later in a tumour's lifetime, and these could lead to treatment resistance.

Tumours contain cells with distinct mutations. Charles Swanton of University College London and his colleagues analysed DNA sequence data from more than 2,500 people with cancer to find out which mutations can fuel the expansion of subpopulations of tumour cells. The team found that mutations that drive cancer growth tend to emerge early in a tumour's evolution, and are typically found throughout the tumour. But the mutations can also develop later, in small subpopulations of cells.

The results suggest that some tumours become resistant to therapy when small populations carrying resistance

genes expand. Understanding which mutations are present in these smaller groups of cells could aid in the selection of therapies for individuals.

Sci. Transl. Med. **7**, 283ra54 (2015)

PALAEONTOLOGY

Ancient seas bore bone-fed worms

Bone-eating worms devour dead whales in today's oceans, but their ancient relatives might have emerged millions of years before their modern food source.

Modern *Osedax* worms drill distinctive holes in bone, with the oldest examples found in whale and fish bones from around 30 million years ago. But Silvia Danise and Nicholas Higgs of Plymouth University, UK, found these boreholes in the fossilized remains of an extinct marine reptile and a turtle, both roughly 100 million years old.

The discovery shifts the evolution of *Osedax*-type worms to well before the rise of early whales, about 50 million years ago. The ancient worms probably survived on the bones of turtles and other animals before whales appeared.

Biol. Lett. **11**, 20150072 (2015)

ECOLOGY

Like a moth to a trumpet flower

Hawk-moths are better at finding nectar in flowers shaped like the bell of a trumpet than in those that resemble a flat disc.

Eric Octavio Campos and his colleagues at the University of Washington in Seattle used a 3D printer to create flowers

that were either flat or curved like a trumpet. The team inserted a tube filled with sugar water into the centre of each flower, and allowed hawk-moths (*Manduca sexta*; **pictured**) to feed from them. Although the moths visited each type of flower equally, they drained trumpet flowers more often than flat ones. The results suggest that the trumpet shape helps to guide the moth's probing mouthparts to nectar.

Before the advent of 3D printers, ecologists who studied plant–pollinator interactions had to breed flowers or sculpt fake ones by hand. The 3D-printing technology allowed Campos's team to quantify flower shape and produce precise models with ease.

Funct. Ecol. **29**, 462–468 (2015)

ANIMAL BEHAVIOUR

Octopus crawls with no rhythm

Octopuses can move quickly in any direction, regardless of which way the eyes and body are facing.

Binyamin Hochner of the Hebrew University in Jerusalem and his colleagues studied the animal's movement by analysing videos of crawling octopuses (*Octopus vulgaris*; **pictured**). They found that the radially symmetrical octopus moves by extending and shortening its arms, and that it tends to crawl diagonally relative to its line of sight.

The octopus also seems to lack any rhythmic pattern when coordinating the movement of its arms, unlike most other animals. The authors suggest that octopuses make last-minute decisions

about which arm to move to determine their direction of crawling.

Curr. Biol. http://doi.org/3vd (2015)

ASTROPHYSICS

Many flavours of supernova

Exploding stars grouped in one family because of their similarities actually form two distinct groups. This may have important cosmic implications because the explosions, called supernovae, are the primary evidence that the Universe's expansion is accelerating.

Half of type Ia supernovae seem to have similar intrinsic brightnesses when seen in the visible spectrum. But when Peter Milne of the University of Arizona in Tucson and his team analysed data from the Hubble Space Telescope and NASA's Swift satellite, they found that the supernovae fell into two subfamilies, each brighter than the other in a different part of the ultraviolet spectrum.

The relative abundances of the two subfamilies seem to have changed over the past several billion years, a fact that could complicate their use as markers of cosmic expansion, the authors say.

Astrophys. J. **803**, 20 (2015)

Full disclosure of clinical trials

Phantom clinical trials are a well-known phenomenon in biomedical research. Trials are conducted, but the results never see the light of day. To tackle this, the World Health Organization (WHO) released a statement in *PLoS Medicine* on 14 April calling for full disclosure of clinical-trial results — a stand that drew wide support from commenters on social media. "Great news!" tweeted Antonietta Filia, an epidemiologist at Italy's National Institute of Health in Rome. Joseph Hayes, a psychiatrist at University College London, suggested in a tweet that such a policy would uncover many less-than-spectacular results. "WHO says all trials should be reported. Journal of Negative & Null Findings to have highest Impact Factor."

↻ **NATURE.COM**
For more on popular papers:
go.nature.com/j5xqms

PLoS Med. **12**, e1001819 (2015)

ANIMAL BEHAVIOUR

Dazzling colours distract predators

Many animals have coloration that shifts depending on the angle from which they are viewed, and this may help them to avoid predators.

This 'interference coloration' has evolved several times in beetles, birds, fish and other creatures, but it is not clear why. Thomas Pike, now at the University of Lincoln, UK, trained laboratory Japanese quails (*Coturnix japonica*) to peck at moving targets on a screen. When the targets changed from green to blue as they moved across the screen, the quails required many more pecking attempts to successfully 'capture' the mock prey. They were also less accurate with their pecks than when the targets stayed the same colour.

The findings suggest that this kind of coloration impairs a predator's ability to accurately locate prey.

Biol. Lett. **11**, 20150159 (2015)

↻ **NATURE.COM**
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

Research boost

In a deal announced on 16 April, Germany's ruling political parties agreed to increase funding for science by €5 billion (US$5.4 billion) between 2018 and 2028. The agreement, which is not yet a government commitment, was welcomed as a strong indication of continuing political support for research. Most of the proposed cash would go towards extending the country's Excellence Initiative, a competitive fund that had been due to expire in 2017. Since its launch in 2006, the initiative has created some 20,000 science jobs. See go.nature.com/ieytqt for more.

Stem-cell initiative

The Center for iPS Cell Research and Application (CiRA) of Kyoto University will partner with Japanese pharmaceutical firm Takeda, in Osaka, to develop clinical applications of induced pluripotent stem (iPS) cells. In a statement on 17 April, Takeda said that it will contribute more than ¥32 billion (US$269 million) to the project over 10 years, as well as to research facilities in Fujisawa. CiRA director Shinya Yamanaka, who shared a Nobel prize in 2012 for his work on iPS cells, will lead the programme. Potential research paths include treatments for heart failure and diabetes.

Ebola-vaccine trial

A late-stage clinical trial of a candidate Ebola vaccine is under way in western Sierra Leone, the US Centers for Disease Control and Prevention announced on 14 April. Conducted in partnership with Sierra Leone's College of Medicine and Allied Health Sciences and its health and sanitation ministry, the trial will enrol 6,000 health and other front-line workers from five heavily stricken districts. Participants will receive the experimental inactivated-virus vaccine rVSV-ZEBOV either immediately, or six months after enrolment. All will be followed closely from the day of enrolment, and those assigned to the later phase will serve as controls for the first phase.

Smoking trends

Electronic cigarettes are now the tobacco products most commonly used by US high-school students. Data released last week by the Centers for Disease Control and Prevention show that a total of 24.6% of students used tobacco products in 2014; 13.4% used e-cigarettes. Declines in conventional cigarette and cigar smoking have been offset by increases in e-cigarettes and hookah use. Mitch Zeller, director of the US Food and Drug Administration's Center for Tobacco Products, called the figures "staggering", and said that they justified his agency's attempts to regulate these products.

Nuclear retractions

Citing a conflict of interests, the journal *Meccanica*, published by Springer, has retracted 11 papers by Alberto Carpinteri, an engineer at the Polytechnic University of Turin, Italy, and his collaborators. Carpinteri was the journal's editor-in-chief until 2014. Some of the papers were about a controversial type of nuclear fission that, according to Carpinteri, is induced by compressing solids (see *Nature* http://doi.org/3wd; 2012). The retractions were issued in March, according to a Springer spokeswoman, but first emerged when a blogger at the Italian newspaper *La Repubblica* spotted them on 15 April.

Telescope share

US astronomers could get more science out of their ground-based telescopes by bartering observation time between facilities, the US National Research Council recommends in a 17 April report. A time-exchange scheme could help researchers to recover from funding cutbacks by the US National Science Foundation, and might assist scientists at small institutions in accessing data from the Large Synoptic Survey Telescope, an 8.36-metre telescope under construction in Chile.

Trial transparency

The World Health Organization issued a statement on 14 April calling

Vertical credit: ESA/ROSETTA/MPS FOR OSIRIS

Dark side of a comet sprouts jet

Pictures taken on 12 March by the European Space Agency's Rosetta spacecraft show the surprise birth of a dust stream from the dark underbelly of the comet 67P/Churyumov–Gerasimenko. The shots were taken just 2 minutes apart, 75 kilometres from the comet's surface.

Jets of vapour and dust normally form at the Sun-facing side of comets as ice sublimates beneath the surface, and researchers had not expected any to emerge from the dark region, named Imhotep after an ancient Egyptian god. But as 67P/Churyumov–Gerasimenko is approaching the Sun, some early light may have warmed an as-yet-unobserved outcrop in the region, the mission team says. The images were presented last week at the annual meeting of the European Geosciences Union in Vienna.

for increased reporting of clinical-trial results. The agency requested that investigators submit their main findings to a peer-reviewed journal and to post key results in clinical-trial registries within one year of completing a study. It also urged researchers to disclose results from older, unreported clinical trials. The statement follows several similar efforts by European and US agencies, and an international campaign by medical researchers (see go.nature.com/p7slov and *Nature* **515,** 477; 2014).

Nuclear veto

A local court in Japan has blocked the restart of two nuclear reactors at the Takahama power plant in the Fukui prefecture. The Nuclear Regulation Authority in Tokyo had deemed the reactors safe to recommence operations last year. But in an injunction issued on 14 April, the presiding judge questioned the stringency of the safety standards, even though the agency had tightened them after the 2011 Fukushima disaster. Despite the national government's efforts to rebuild the nuclear industry, all of Japan's reactors remain offline.

Porpoise protection

On 16 April, Mexican President Enrique Peña Nieto officially launched a plan to save the vaquita (**pictured**), a critically

endangered porpoise that lives only in the Gulf of California. Fewer than 100 vaquitas (*Phocoena sinus*) are estimated to remain. The animals often become ensnared in gillnets — a type of fishing net. The US$37-million initiative prohibits gillnet fishing in much of the northern gulf, supplies new boats to the Mexican navy to enforce the ban, and compensates fishers for lost income.

Corneal implant

The US Food and Drug Administration approved a first-of-its-kind corneal implant on 17 April. The KAMRA inlay is an opaque, ring-shaped device to improve close-up vision in people with presbyopia, an age-related condition that diminishes the eye's ability to change focus. The implant blocks peripheral light and funnels central light rays through a small opening, sharpening vision of near objects and small print. Made by AcuFocus, a medical-device

company in Irvine, California, the implant is approved for people who have not had cataract surgery.

Telescope kick-off

Construction of the Large Synoptic Survey Telescope officially launched on 14 April with a traditional Chilean ceremony to lay the first stone on Cerro Pachón, a mountain in north-central Chile. The 8.36-metre telescope, equipped with a 3-billion-pixel digital camera, is expected to come online in 2019 and begin full operation in 2022. The construction, projected to cost US$680-million, is funded by the US National Science Foundation, the US Department of Energy and private donors.

Trachea inquiry

An investigation by the Karolinska Institute in Stockholm has cleared one of its surgeons, Paolo Macchiarini,

of charges of scientific misconduct related to a series of synthetic-trachea transplants that were performed starting in 2008. Pierre Delaere, a surgeon at the hospital of the Catholic University of Leuven (KU Leuven) in Belgium, had accused Macchiarini of overstating the success of the operations he had performed, among other allegations (see *Nature* **516,** 16–17; 2014). The investigation report by the Karolinska ethics council, released on 14 April, judged Macchiarini's claims to be plausible. Another Karolinska investigation, which will compare patient medical records with Macchiarini's claims, is ongoing.

Chemical weapons

One hundred years have passed since the first large-scale use of chemical weapons at Ypres, Belgium, on 22 April 1915, during the First World War. The European Association for Chemical and Molecular Sciences is marking the anniversary by calling for the complete elimination of chemical weapons; the society, which has members from 32 countries in Europe, has organized a three-day conference in Ypres to discuss how to improve international laws to prevent the use of such weapons.
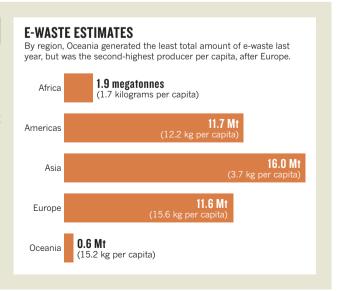
26–30 APRIL
The first Triennial Earth–Sun Summit takes place in Indianapolis, Indiana. Heliophysicists discuss topics such as space weather, solar flares and results from NASA's MAVEN mission to explore how Mars's atmosphere interacts with solar wind.
go.nature.com/fc73ho

TREND WATCH

Of the 41.8 million tonnes of electronic and electrical waste that was generated in 2014, nearly 60% was a mix of kitchen, bathroom and laundry equipment, according to a report released on 19 April by the United Nations University in Bonn, Germany (see go.nature.com/hznxhn). About 7% consisted of personal computers and other small information-technology products. Globally, an average of 5.9 kilograms of e-waste was generated per capita — a figure that the authors project will rise to 6.7 kg in 2018.

E-WASTE ESTIMATES

By region, Oceania generated the least total amount of e-waste last year, but was the second-highest producer per capita, after Europe.

Africa — **1.9 megatonnes** (1.7 kilograms per capita)

Americas — **11.7 Mt** (12.2 kg per capita)

Asia — **16.0 Mt** (3.7 kg per capita)

Europe — **11.6 Mt** (15.6 kg per capita)

Oceania — **0.6 Mt** (15.2 kg per capita)

↻ **NATURE.COM**
For daily news updates see:
www.nature.com/news

Grey wolves are increasingly rare on Isle Royale.

ECOLOGY

Iconic island study at risk

Just three wolves left in 57-year ecology project.

BY EMMA MARRIS

The longest-running predator–prey study in the world may run out of predators as early as next year. Isle Royale in Michigan is down to its last three wolves, researchers announced on 17 April.

That spells the probable end of a 57-year project to chronicle the fluctuating populations of moose and wolves on the 544-square-kilometre island. The effort has tested mathematical models of population dynamics, and provided insight into wolf and moose behaviour.

But inbreeding has taken its toll on the wolf population, causing skeletal abnormalities and high death rates; the last influx of DNA came in 1997, when a lone male travelled to the island in Lake Superior over a temporary ice bridge into the island, which is a national park. (By contrast, the moose population has grown by about 22% per year, from some 500 in 2011 to 1,250 now.) For years, project leaders Rolf Peterson and John Vucetich, both ecologists at Michigan Technological University in Houghton, called for a 'genetic rescue' to save the study. They say that by transporting just a few wolves onto the island, problems such as malformed spinal columns, caused by recessive genes, would be masked by the incoming dominant genes.

This year, Peterson and Vucetich flew above the island for 16 days in February and March looking for wolves and wolf tracks in the snow. Most days they came up with nothing. "Just a blank slate," says Peterson. In the end, they counted just three wolves — probably a mated pair of close adult relatives and one pup.

Even if wolves were brought in this year, the pair probably would not break up and mate with the newcomers, the researchers say. And the pup was in such poor shape that the scientists do not expect it to survive. "Genetic rescue is almost certainly too late now," says Vucetich.

Peterson blames the US National Park Service (NPS) for "dithering" over what to do until the window for genetic rescue had closed. A year ago, the NPS announced that it would make a decision on the matter after a "comprehensive planning process".

"We have science coming out our ears and it wasn't enough to carry the day," says Peterson.

MIXED FEELINGS

But Phyllis Green, superintendent of Isle Royale National Park, says that science is just one of several factors that will influence her decision about whether to intervene, along with NPS policies, relevant laws and the needs and desires of park visitors. "Some people love hearing a wolf howl in the wilderness; others say 'if you put the wolf out there, I am not as thrilled about hearing it'," she says.

And not everyone agrees that the NPS should have added more wolves to the island. Prominent wolf biologist David Mech, who works for the US Geological Survey in St Paul, Minnesota, says that much can be learned from watching what happens without intervention. "Chances are good that one [of the remaining wolves] is a male and one is a female, and that is how the whole population started, so it would be interesting to see what happens in the next year," he says. "I think we should just continue to study the situation and describe what happens."

In March, the US National Science Foundation renewed the grant for the Isle Royale study for another five years, and that US$90,000 is not contingent on wolves remaining in the system. Moose and vegetation can still be tracked as they react to the disappearance of the predator. The researchers say that they will stay. Vucetich has been working on the study since the early 1990s, and Peterson has been looking for wolf tracks in the snow every winter since 1971. "They can kick me out if they want," Peterson says, "but I won't walk away." ∎

Photo credit: JOHN VUCETICH/MICHIGAN TECHNOLOGICAL UNIV./AP

A study in Sweden monitored how bees respond to neonicotinoids in the wild.

MAJ RUNDLÖF

Bee studies stir up pesticide debate

The threat that neonicotinoids pose to bees becomes clearer.

BY DANIEL CRESSEY

The case for restricting a controversial family of insecticides is growing. Two studies published on 22 April in *Nature*[1,2] address outstanding questions about the threat that the chemicals pose to bees, and come as regulators around the world gear up for a fresh debate on pesticide restrictions.

Many bee populations are in steep decline, with multiple causes identified, including parasites and the loss of food sources. Also blamed are neonicotinoids, a widely used class of insecticides that are often applied to seeds, and find their way into the pollen and nectar of plants. The use on seeds of three — clothianidin, imidacloprid and thiamethoxam — is temporarily banned in the European Union because of concern that they might harm pollinators; the ban is up for review in December. In the United States, there are no such restrictions, but the US Environmental Protection Agency said on 2 April that it was "unlikely" to approve new outdoor neonicotinoid-pesticide uses without new bee data.

So far, the data are mixed. Many studies that link the poor health of bee colonies to the pesticides have been criticized, for example for not using realistic doses. Some defenders of the chemicals have argued that if neonicotinoids are harmful, bees will learn to avoid treated plants.

Geraldine Wright, an insect neuroethologist at Newcastle University, UK, and her colleagues investigated this aspect. They confined honeybees (*Apis mellifera*) and bumblebees (*Bombus terrestris*) to boxes and gave them a choice between plain nectar and nectar laced with imidacloprid, thiamethoxam or clothianidin. The researchers found that the bees showed no preference for the plain nectar. In fact, the insects were more likely to choose the nectar containing imidacloprid or thiamethoxam[1], although it is not clear whether the preference would occur in the wild.

Wright's team also analysed the response of the bees' taste neurons to neonicotinoids, and found that they reacted the same regardless of concentration — indicating that the bees cannot taste the pesticides and that the preference is caused by some other mechanism. Other studies have shown that neonicotinoids activate receptors in bee brains linked to memory and learning.

In contrast to Wright and colleagues' work, the second paper[2] looked at honeybees and wild bees, including bumblebees, in the field. Maj Rundlöf, an ecologist at Lund University in Sweden, and her colleagues analysed eight fields of oilseed rape sown with seeds treated with clothianidin and eight fields sown with untreated seeds across southern Sweden.

Honeybees did not respond differently in the treated and untreated fields. But the researchers found that wild-bee density in treated fields was around half that in untreated fields. Nests of solitary bees and bumblebee-colony growth were also reduced in treated fields. "I'm worried about the effects on wild bees," says Rundlöf.

She suggests that honeybees have larger colony sizes, which could sustain higher losses of foraging bees before showing overall health effects. But that suggests another problem. "Honeybees are the model organism that is used in toxicity testing for pesticides," she says. If they are not representative of bees in general, it could explain why more studies have not detected negative effects.

Dave Goulson, a bee researcher at the University of Sussex in Brighton, UK, also suspects that honeybees are more resilient than wild bees to neonicotinoids. Rundlöf's paper is "probably the best field study done so far", he says, and avoids many previous problems, such as contaminated controls. "Any reasonable person would have to accept this is a real effect," he adds.

The debate is heating up. In March, Goulson reanalysed[3] data from a 2013 study by the UK Food and Environment Research Agency (see go.nature.com/w9jlti), which had concluded that neonicotinoid pesticides do not harm bees: Goulson found that they do. In the same month, work from the United States found[4] that the probable harm from exposure to imidacloprid in seed-treated crops was "negligible" in honeybees, and last year a study[5] done in Canada reached a similar conclusion for clothianidin on oilseed rape.

Christopher Connolly, who studies human and bee neuroscience at the University of Dundee, UK, and has published work[6] showing that neonicotinoids interfere with neuron function in bumblebees, says that he was already convinced that the pesticides are bad for bees. Now, "the questions need to move to a different level", to elucidate the mechanisms. ∎

1. Kessler, S. C. *et al. Nature* http://dx.doi. org/10.1038/nature14414 (2015).
2. Rundlöf, M. *et al. Nature* http://dx.doi.org/10.1038/ nature14420 (2015).
3. Goulson, D. *PeerJ* **3**, e854 (2015).
4. Dively, G. P., Embrey, M. S., Kamel, A., Hawthorne, D. J. & Pettis, J. S. *PLoS ONE* http://dx.doi.org/10.1371/ journal.pone.0118748 (2015).
5. Cutler, G. C., Scott-Dupree, C. D., Sultan, M., McFarlane, A. D. & Brewer, L. *PeerJ* http://dx.doi. org/10.7717/peerj.652 (2015).
6. Moffat, C. *et al. FASEB J.* http://dx.doi.org/10.1096/ fj.14-267179 (2015).

NEUROLOGY

Drug that boosts nerve signals offers hope for multiple sclerosis

Trialled antibody treatment thought to work by renewing the protective coating of neurons.
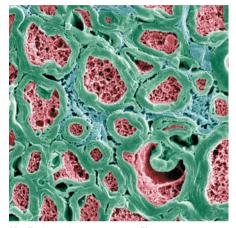
BY HEIDI LEDFORD

An experimental antibody drug aimed at protecting nerves from the ravages of multiple sclerosis offers hope for a new way to combat the neurological disease — if researchers can definitively show that it works.

The antibody, anti-LINGO-1, is intended to stimulate regrowth of the myelin sheath, the fatty protective covering on nerve cells that is damaged by multiple sclerosis. Its developer, Biogen of Cambridge, Massachusetts, will present results from a small clinical trial at an American Academy of Neurology meeting this week in Washington DC. If the initial promising results from the trial are confirmed, it will be the first such myelin-regeneration therapy. Other researchers are racing to find more targets and compounds that act similarly.

"Once we get a positive result, the field will move very quickly," says Jack Antel, a neurologist at McGill University in Montreal, Canada. But that excitement is tempered by practical hurdles: there is as yet no proven way to measure remyelination of nerve cells in living humans.

Myelin sheaths insulate and support axons, the fibres that transmit signals between nerve cells. In multiple sclerosis, immune attack destroys these sheaths. Stripped of this protective coating, the axons gradually wither away, causing the numbness and muscle spasms that are characteristic of the disease. The 12 drugs approved in the United States to treat multiple sclerosis slow this immune attack — although sometimes with dangerous side effects. But none stops it, says Bruce Trapp, a neuroscientist at the Cleveland Clinic in Ohio.

Anti-LINGO-1 blocks the LINGO-1 protein, which inhibits the production of myelin. In doing so, the drug spurs myelin growth. It has consistently performed well in animal models and in human cells grown in culture.

Myelin (green) protects nerve fibres.

LINGO-1 is not the only target for myelin-boosting therapies. Acorda Therapeutics of Ardsley, New York, is conducting clinical trials of an antibody that binds to the cells that give rise to myelin, although its molecular target is unknown. And this week in *Nature*, researchers report that two drugs marketed for skin conditions help to repair myelin in mice and in cultured human cells (F. J. Najm *et al. Nature* http://dx.doi.org/10.1038/nature14335; 2015).

Numerous studies have also found early drug leads, some of them among previously approved therapies. "We're entering the golden age of targets," says Robert Miller, a neuroscientist at George Washington University in Washington DC, who collaborates with Biogen.

Biogen is in the lead, and the industry is watching to see how the company will ascertain whether its drug is working, says Johanne Kaplan, vice-president of neuroimmunology research at Genzyme, a company in Cambridge, Massachusetts, that is also developing remyelinating drugs. Assessing multiple-sclerosis drugs is challenging: symptoms are difficult to measure reliably and often progress slowly. And the benefits of remyelinating drugs are expected to manifest over years, not months.

At the neurology meeting, Biogen will report on a trial of anti-LINGO-1 in 82 patients with optic neuritis, a loss of vision common among people with multiple sclerosis. The trial found that the antibody was no better than a placebo at improving vision. But it did speed up signalling in retinal nerves — a possible sign that the myelin sheath had been rebuilt — by 41% over placebo after eight months of treatment, says Gilmore O'Neill, vice-president of multiple-sclerosis research at Biogen.

The trial was small, and it is too early to know whether the drug will provide tangible relief from other symptoms of multiple sclerosis, such as numbness, cautions Charles ffrench-Constant, who studies the disease at the University of Edinburgh, UK. Similar improvements in nerve conductivity could result from reduced inflammation, he adds. "You can't assume that this is actually evidence of myelin regeneration."

Nevertheless, ffrench-Constant and others are excited to see any sign that the approach may be working. "It is remarkable," says Vittorio Gallo, a neuroscientist at the Children's Research Institute in Washington DC. Biogen is conducting another anti-LINGO-1 trial to test an experimental way (based on magnetic resonance imaging) to assess remyelination, as well as the effect of the drug on symptoms in people for whom the disease is progressing rapidly.

There is little expectation that anti-LINGO-1 or other drugs like it will cure multiple sclerosis, says Kaplan. Instead, the hope is that the drugs can be used in combination with the available immune treatments to slow progression of the disease. "If we could get the therapies to patients early enough, maybe we can keep them out of the wheelchair," says Trapp. "That would be huge." ∎

STEVE GSCHMEISSNER/SPL/CORBIS

SILULATIONS: S. SHAPIRO, R. GOLD, V. PASCHALIDIS, M. RUIZ, Z. ETIENNE & H. PFEIFER. MOVIE: S. SHAPIRO, S. CONNELLY, A. KHAN & L. KONG.

MORE ONLINE

VIDEO

3D models of colliding black holes are most realistic yet go.nature.com/5vxp8t

MORE NEWS

● US societies push back against NIH reproducibility rules go.nature.com/f7col7
● Shipwrecked 170-year-old champagne gives clues to winemaking's past go.nature.com/cqdwja
● Foodies embrace 3D-printed cuisine go.nature.com/8xg1zp

TOP STORY

Homo erectus footprints hint at ancient hunting party go.nature.com/glmcnk

BRIAN RICHMOND

Seismologist Amberlee Darold is tracking unprecedented earthquakes in Oklahoma, and trying to discover what is causing them.

GEOLOGY

Artificial quakes shake Oklahoma

Earthquakes linked to oil and gas operations prompt further research into human–induced seismic hazards.

BY ALEXANDRA WITZE, STILLWATER, OKLAHOMA

It's the first thing that geologist Todd Halihan asks on a sunny spring afternoon at Oklahoma State University in Stillwater: "Did you feel the earthquake? My mother-in-law just called to complain that the house was shaking."

Halihan's mother-in-law has been calling a lot lately. Fifteen quakes of magnitude 4 or greater struck in 2014 — packing more than a century's worth of normal seismic activity for the state into a single year. Oklahoma had twice as many earthquakes last year as California — a seismic hotspot — and researchers are racing to understand why before the next major one strikes.

Whatever they learn will apply to seismic hazards worldwide. Oklahoma's quakes have been linked to underground wells where oil and gas operations dispose of waste water,

but mining, geothermal energy and other underground explorations have triggered earthquakes from South Africa to Switzerland. This week, at a meeting of the Seismological Society of America in Pasadena, California, scientists will discuss how the risk from human-induced quakes differs from that of natural quakes — and how society can prepare for it.

In Oklahoma, the earthquakes have unleashed a frenzy of finger-pointing, with angry residents suing oil and gas companies over damage to their homes. The industry and politicians are locked in fierce debates about whether the quakes are induced, but the unprecedented shaking across central and northern parts of the state matches almost exactly with the activity of water-disposal wells. "There are some who will argue that it is purely natural," says Halihan. "But by now it's pretty clear it's not."

Companies drill into the ground to extract oil and gas mixed with salt water, essentially

the brine from a long-fossilized sea. They separate out the fuels and then inject the salt water into deep disposal wells (there are more than 4,600 in Oklahoma). State regulations require that the salt water be disposed of in rock layers below those that hold drinking water (see 'Shaky ground').

STRESS FRACTURE

Much of the liquid ends up in a rock formation called the Arbuckle, which underlies much of Oklahoma and is known for its ability to absorb huge volumes of water. But in many places the Arbuckle rests on brittle, ancient basement rocks, which can fracture along major faults under stress. "The deeper you inject, the more likely it is that the injected brine is going to make its way into a seismogenic fault zone, prone to producing earthquakes," says Arthur McGarr, who leads research on induced quakes at the US Geological Survey (USGS) in Menlo Park, California.

Oil and gas companies operate disposal wells across the central United States, and although Oklahoma stands out for the sheer volume of waste water, other states may be getting triggered earthquakes. A report in *Nature Communications*[1] this week, for example, links brine injection to a series of quakes that began in November 2013 near Azle, Texas.

The basic physics of the process has been understood since the 1970s, when scientists from the USGS pumped water down a well in Rangely, Colorado, and recorded how earthquake activity rose and petered out as they varied the amount of fluid[2]. The question now is which faults are likely to rupture in

Oklahoma, and how large an earthquake they might produce.

Whether a fault breaks in an earthquake depends on how it sits in relation to the stresses that compress Earth's crust. The movement of tectonic plates is squeezing Oklahoma from east to west, so most of the earthquakes are happening along faults oriented northwest to southeast, or northeast to southwest. Other faults are less likely to rupture, says McGarr.

The biggest earthquake ever recorded in Oklahoma was a magnitude-5.6 event near the town of Prague in November 2011, and many seismologists think that it was induced by nearby disposal wells[3]. Theoretical work[4] suggests that the potential size of a quake grows with the volume of fluid injected into the ground. The biggest disposal wells in Oklahoma inject more than 60 million litres of waste water each month.

Austin Holland, the state seismologist at the Oklahoma Geological Survey in Norman, estimates that the chance of another earthquake of magnitude 5 or greater striking the state in the next year is about 30%. "That is not the kind of lottery we want to win," he says.

Oklahoma has designated buffer zones, requiring extra scrutiny for disposal wells within 10 kilometres of sites of earthquake swarms or quakes of magnitude 4 or greater. As of 18 April, operators must also prove they are not injecting into or near basement rocks, or must cut their disposal volumes by half.

Yet oil and gas companies hold great political power in Oklahoma, and regulators continue to emphasize what they call uncertainty in linking injection wells to quakes. "We felt a big quake one Friday night and I knew we had permitted a brand-new Arbuckle disposal well not three miles from my house," said Tim Baker, director of the oil and gas division of the Oklahoma Corporation Commission, which regulates drilling, at a town-hall meeting in suburban Oklahoma City this month. "I drove to that well to inspect it on Saturday morning, and it wasn't even turned on. That's how complex this issue is."
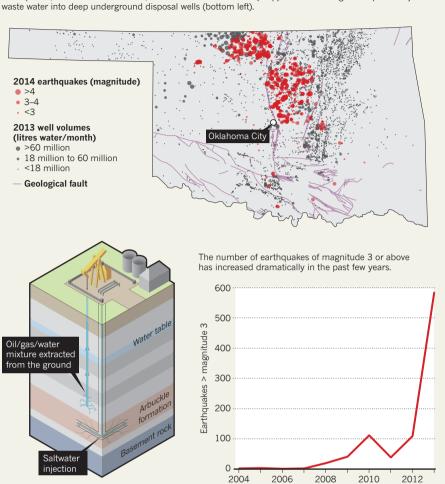
The related — and controversial — technique of hydraulic fracturing, in which water is injected into rock to open cracks so oil and gas can flow more easily, has also been linked to earthquakes, but to a much lesser extent. The fracking involves injecting less water for shorter periods of time, and has not been tied to any earthquakes greater than magnitude 4 (ref. 5).

SEISMIC SURVEY

One group of geologists wants to explore exactly how disposal wells might cause earthquakes. The team hopes to find a remote corner of Oklahoma and inject fluids deep underground while monitoring seismicity, in a modern analogue to the 1970s experiments in Colorado. "It's a very ambitious goal, but we want to do a controlled field-scale experiment," says Ze'ev Reches, a geophysicist at the University of Oklahoma

SHAKY GROUND

Earthquakes, once rare in Oklahoma, are now common in areas (map) where oil and gas companies inject waste water into deep underground disposal wells (bottom left).

2014 earthquakes (magnitude)
● >4
● 3–4
· <3

2013 well volumes (litres water/month)
● >60 million
● 18 million to 60 million
· <18 million

— Geological fault

Oklahoma City

Oil/gas/water mixture extracted from the ground

Water table

Arbuckle formation

Basement rock

Saltwater injection

The number of earthquakes of magnitude 3 or above has increased dramatically in the past few years.

in Norman and a co-leader of the project. But with Oklahomans already on edge, it is not clear whether the team could pull off such an experiment. So far, it remains hypothetical.

For now, seismologists are just trying to keep up with the quakes. The state geological survey recently gave up naming earthquake swarms, because the quakes simply never stopped, says Amberlee Darold, an agency seismologist. (The survey used to name swarms after nearby towns; it now identifies huge swathes of continuous activity by county.)

In the 15-storey brick Earth sciences building on the University of Oklahoma campus in Norman, statues celebrate the state's 'wildcatters' who made it big in oil and gas, and a well-manicured garden nearby is dedicated to their achievements. Holland and Darold labour in the building's dark basement, compiling a database of Oklahoma's faults and trying to make sure that every earthquake is documented.
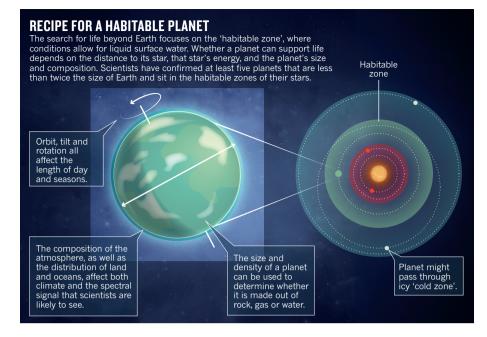
Many scientists are worried that the state's buildings are not constructed to standards that consider seismic risk, and are concerned about how old brick-and-mortar structures would hold up in a large earthquake. The USGS issues national seismic-hazard maps every few years,

but has never included the risk from induced quakes. This year, for the first time, the agency is developing induced-seismicity hazard maps for Oklahoma and surrounding states. The first of these is likely to be out by the end of 2015, says McGarr.

In Cushing, almost 60 kilometres north of Prague, crude-oil pipelines from across the continent meet. Fences topped with razor wire are meant to protect huge oil-storage tanks from a terrorist attack, but will not help if a major earthquake strikes, says Halihan.

In the meantime, he sits and waits to hear about the next quake. If he does not want to rely on his mother-in-law, Halihan can track the tremors by watching the movement of a small brass marker pinned to his office wall. It used to shake about once a week. Now it does so almost every day. ■

1. Hornbach, M. J. et al. Nature Commun. **6,** 6728 (2015).
2. Raleigh, C. B., Healy, J. H. & Bredehoeft, J. D. Science **191,** 1230–1237 (1976).
3. Keranen, K. M., Savage, H. M., Abers, G. A. & Cochran, E. S. Geology **41,** 699–702 (2013).
4. McGarr, A. J. Geophys. Res. Solid Earth **119,** 1008–1019 (2014).
5. Ellsworth, W. L. Science **341,** 1225942 (2013).

RECIPE FOR A HABITABLE PLANET

The search for life beyond Earth focuses on the 'habitable zone', where conditions allow for liquid surface water. Whether a planet can support life depends on the distance to its star, that star's energy, and the planet's size and composition. Scientists have confirmed at least five planets that are less than twice the size of Earth and sit in the habitable zones of their stars.

Orbit, tilt and rotation all affect the length of day and seasons.

The composition of the atmosphere, as well as the distribution of land and oceans, affect both climate and the spectral signal that scientists are likely to see.

The size and density of a planet can be used to determine whether it is made out of rock, gas or water.

Habitable zone

Planet might pass through icy 'cold zone'.

PLANETARY SCIENCE

Climate scientists eye alien worlds

NASA initiative seeks to bolster interdisciplinary science in the search for life beyond Earth.

BY JEFF TOLLEFSON

The hunt for life beyond the Solar System is gaining new partners: NASA climatologists. After more than 30 years of studying Earth, a team at the NASA Goddard Institute for Space Studies (GISS) in New York will adapt its global climate model to simulate conditions on potentially habitable exoplanets. The effort is part of a broader push to identify Earth-like worlds that NASA will launch on 20 April at a meeting in Washington DC.

Already, the agency's space-based Kepler telescope has pinpointed more than 1,000 alien planets by observing the brief interruption of starlight that signals a planet passing in front of its parent star. At least five of these planets are similar in size to Earth and located in the 'habitable zone', where liquid water could persist. The next step would be to detect light passing through exoplanet atmospheres, which could hold clues to conditions on these distant worlds.

"We have to start thinking about these things as more than planetary objects," says Anthony Del Genio, a climate modeller who is leading the GISS effort. "All of a sudden, this

has become a topic not just for astronomers, but for planetary scientists and now climate scientists."

Del Genio's group is one of around 16 — ranging from Earth and planetary scientists to solar physicists and astrophysicists — that are participating in NASA's new Nexus for Exoplanet System Science (NExSS) programme. The effort has an initial annual budget of roughly US$10 million to $12 million.

"We are bringing together a bunch of different disciplines, and they all look at the formation and functioning of planets in different ways," says Mary Voytek, who directs NASA's astrobiology programme and organized NExSS. Although interest is high, Voytek says, communication remains a challenge. "You can't even get these communities to agree on a definition of the habitable zone."

The initiative is based in part on the Virtual Planetary Laboratory at the University of Washington in Seattle, which launched in 2001 and now has 55 researchers from 23 institutions collaborating on interdisciplinary exoplanet research. "When we started this, people thought we were crazy," says Victoria Meadows, an astronomer and

the project's principal investigator. "But this is not something that a single discipline can address."

NExSS will expand the network of researchers collaborating on exoplanets, she says. That should help scientists to make sense of existing data and observations from the James Webb Space Telescope and the Transiting Exoplanet Survey Satellite, which are both scheduled for launch in 2018. It could also help NASA develop missions to hunt for exoplanets in the 2020s and beyond.

At GISS, Del Genio's team has already started repurposing the institution's workhorse Earth-system climate model. The researchers are combing through its source code to locate simple parameters that are fixed for Earth, such as 24-hour days and 365-day orbits, in order to create an exoplanet model that can be adjusted for different planetary systems. Initial simulations will focus on Earth's ancient past and the evolution of Mars and Venus. Although neither can support life today, each may have had liquid surface water at some time.

The GISS team's ultimate goal is to explore the concept of a habitable zone by mixing and matching some of the key factors that determine whether a planet can support life (see 'Recipe for a habitable planet'). By feeding these parameters into the exoplanet model, the group will create a database of 'hypothetical atmospheres' with spectra that could be visible to astronomers.

Del Genio's group is one of several that are using climate models in exoplanet research. A group led by physicist François Forget at the Pierre Simon Laplace Institute in Paris used such a model to explore the runaway warming that occurs when water vapour builds up in a planet's atmosphere, trapping outgoing radiation. In December 2013, they reported in *Nature*[1] that the early Earth could withstand more solar radiation before its oceans boiled off than scientists had thought. Two other groups have done similar experiments[2,3], and Del Genio says that his team is exploring the same issue.

The 3D models could be particularly useful for defining the habitable zone, and the place to start is the Solar System, says James Kasting, an atmospheric scientist at Pennsylvania State University in University Park who works on one-dimensional exoplanet models. But ultimately, he adds, progress will depend on better observations of exoplanets.

"In 15 or 20 years, we might get a spectrum of a planet that looks Earth-like, and then everyone will be out with their models trying to model that planet," he says. "I would like it to happen quicker — but we need a big telescope." ∎

1. Leconte, J. *et al. Nature* **504,** 268–271 (2013).
2. Yang, J., Cowan, N. B. & Abbot, D. S. *Astrophys. J.* **771,** L45 (2013).
3. Wolf, E. T. & Toon, O. B. *Geophys. Res. Lett.* **41,** 167–172 (2014).

PALAEOANTHROPOLOGY

Oldest stone tools raise questions about their creators

The 3.3-million-year-old implements predate the first members of the Homo *genus.*

BY EWEN CALLAWAY, SAN FRANCISCO

The oldest stone tools on record may spell the end for the theory that complex toolmaking began with the genus *Homo*, to which humans belong. The 3.3-million-year-old artefacts, revealed at a conference in California last week, predate the first members of *Homo*, and suggest that more-ancient hominin ancestors had the intelligence and dexterity to craft sophisticated tools.

"This is a landmark discovery pertaining to one of the key evolutionary milestones," says Zeresenay Alemseged, a palaeoanthropologist at the California Academy of Sciences in San Francisco, who attended the talk at the annual meeting of the Paleoanthropology Society in San Francisco, on 14 April.

More than 80 years ago, anthropologist Louis Leakey found stone tools in Olduvai Gorge in Tanzania. Decades later, he and his wife Mary and their team found bones from a species that the Leakeys named *Homo habilis* — 'the handy man'. This led to the prevailing view that human stone-tool use began with *Homo*, a group that includes modern humans and their big-brained and tall forebears. The oldest of these Oldowan tools date to 2.6 million years ago — around the time of the earliest *Homo* fossils. Climate upheavals that transformed dense forest into open savannah might have catalysed ancient humans into developing the new technology so that they could hunt or scavenge grass-eating animals, the theory goes.

Chimpanzees and other non-human primates use stones to crack nuts, for instance, but their tools lack the craftsmanship of the Oldowan toolmakers, who would strike one rock against another, breaking off flakes to leave a sharp-edged stone core.

In 2010, Alemseged and his team reported an intriguing find at a site called Dikika in Ethiopia (S. P. McPherron *et al. Nature* **466**, 857–860; 2010). They saw cut marks on bones from 3.4 million years ago, when ape-like creatures such as *Australopithecus afarensis* — the same species as the famous fossil called Lucy — roamed eastern Africa. This hinted at even earlier manufacturing of stone tools. Other researchers questioned the find, attributing the marks to natural wear and tear such as trampling, or bites inflicted by crocodiles.

Aware of this controversy, a team led by Sonia Harmand of Stony Brook University in

Excavators at Lomekwi, Kenya, in 2011.

New York set out in 2011 to find tools older than 3 million years, at a site west of Kenya's Lake Turkana. On a July day, the team took a wrong turn and happened upon a patch of land that seemed worth exploring. By tea time, they had found pieces of rock lying on the ground that looked like flakes left over from the manufacture of stone tools. Careful excavation of the patch revealed 19 buried artefacts, including stone core forms, and dozens more on the surface. One key surface find was a small rock flake, which fitted in a gap in a buried core as snugly as a jigsaw puzzle piece, confirming that the tools were made through a flaking process.

The tools come from sediments that Harmand's team dated to around 3.3 million years ago and are much larger than the Oldowan artefacts: some weigh as much as 15 kilograms. The team concluded that the tools represent a distinct culture, which they have named the Lomekwian culture after the site where the implements were found. "Lomekwi marks a new beginning to the known archaeological record," Harmand said at the meeting.

Hominin fossils and cut-marked animal bones have not been found at the site, so the team cannot yet say who made the tools or how they were used. But their discovery may deliver a fatal blow to the already fragile idea that complex toolmaking began with *Homo*. Harmand suggests that earlier species, such as *Kenyanthropus platyops*, bones of which have been found on the western shore of Lake Turkana, and *A. afarensis*, may have made tools by building on the cruder abilities seen in apes and monkeys. The Lomekwi tools were made in a forest environment, also questioning the idea that open landscapes catalysed tool use, said Harmand.

Alemseged sees the Lomekwi tools as vindication for his team's controversial find of cut-marked bones. Before Harmand's presentation, Alemseged's colleague Jessica Thompson, an archaeologist at Emory University in Atlanta, Georgia, presented an analysis of other animal bones from Dikika. None contained similar patterns to those reported in 2010, suggesting that the marks were made by something other than wear and tear — probably by tools.

The Lomekwi talk left David Braun, an archaeologist at George Washington University in Washington DC, itching for further details. He says that the tools look authentic, as does the date that Harmand and her team assert. The identity of their makers has aroused his curiosity: "What the hell do these things look like if they can use 15-kilogram tools?"

But he is most interested in what the Lomekwi tools meant for their creators. Did they offer an advantage over the other hominins that were around at the time, or was toolmaking more common 3 million to 4 million years ago than existing evidence suggests? "They're a game-changer," he adds, "no matter what." ∎

CORRECTIONS

The News story 'Hope for science in fallout of nuclear deal' (*Nature* **520,** 274–275; 2015) wrongly stated that Iran found a bank willing to accept its payment of dues to CERN. It was CERN, not Iran, that found the bank. In addition, the text implied that Iran is a full member of CERN — it is involved in specific projects at the lab. And the News story 'Leading scientists favour women in tenure-track hiring test' (*Nature* **520,** 275–276; 2015) misidentified Virginia Valian as a linguist. She is a psychologist. It also misrepresented her views on the study by Williams & Ceci: she has reservations about aspects of the study's methodology.

The soil sleuth

Forensic geologist Lorna Dawson has pioneered methods to help convict criminals using the dirt from their shoes.

BY CHELSEA WALD

On a Saturday night in October 1977, Lorna Dawson was studying in her dormitory at the University of Edinburgh, UK, when two 17-year-old girls disappeared off a nearby street. The teenagers had been on a pub crawl with friends, stopping at an old Scottish tavern called The World's End before vanishing. Police officers remembered seeing two men with them. The next day, the girls turned up dead eight kilometres apart — one on a beach and another in a remote wheat field. They had both been raped, beaten and strangled. Despite a nationwide manhunt, police could not find the assailants.

Dawson was a country girl, new to the city, and was working towards a geology degree at the time of the crime, later dubbed the World's End murders. "It was my first time away from home," she says, and the case left her "terrified to go out".

It also left her with a passion for justice. Now at the James Hutton Institute in Aberdeen, UK, Dawson runs one of the world's only labs dedicated to forensic soil science, where in the past decade she has worked on more than 70 cases from around the globe. At the time of the murders, soil was rarely used as evidence, and techniques were "elementary", she says. But today, soil evidence regularly leads to bodies, overturns alibis and reveals the origins of artefacts. That is in no small part due to Dawson, who has advanced methods in soil forensics and worked to disseminate the techniques to others.

"Lorna has been instrumental in promoting the new renaissance in forensic geoscience throughout the world," says Marianne Stam, who recently retired from the Riverside Crime Laboratory of the California Department of Justice. But Dawson says that there is still more to be done. She is now part of an international collaboration developing a method to profile microbial communities using DNA. This could make soil more valuable for crime fighting, says Rob Fitzpatrick, who founded the Centre for Australian Forensic Soil Science in Adelaide, Australia. "What Lorna is doing is pioneering new ground, developing methods that others could use and should try more."

Dawson says that she does it for the victims, such as the girls in the World's End murders — a case she would return to several times in her career. "Getting some sort of closure for the victims' families, it's a really rewarding thing to be able to do," she says. "That's what drives you on into the hard hours."

BEYOND SHERLOCK HOLMES

Forensic soil science was nearly 150 years old by the time Dawson took it up. Police in Germany used sandy soil to solve a crime in 1856 and Arthur Conan Doyle noted the forensic potential of dirt three decades later in his first Sherlock Holmes mystery, *A Study in Scarlet*. In the story, Watson says that Sherlock "tells at a glance different soils from each other. After

walks, has shown me splashes upon his trousers and told me by their colour and consistence in what part of London he had received them."

In the real world, forensic soil science advanced little beyond analyses of "colour and consistence" for the next century. "It was really just about the larger components of the soil. You took a sample and you shook it up in a test tube and noted its colour, that sort of thing," says retired forensic scientist Dave

> "GETTING SOME SORT OF CLOSURE FOR THE VICTIMS' FAMILIES, IT'S A REALLY REWARDING THING TO BE ABLE TO DO."

Barclay. From 1997 to 2006, he was head of physical evidence at the National Crime Faculty, a forensics agency that helped UK police departments with the most serious crimes. When he looked at the state of soil forensics, he was dismayed. "It wasn't being used, or it was being done on a sort of cottage-industry basis," he says. "It wasn't being done necessarily to the standards of normal science."

A few scientists were starting to rigorously apply new techniques, such as specialized scanning electron microscopy for mineralogy, but other practitioners continued to use unsophisticated methods to give investigators the answers they wanted, he says. In other countries, the situation was similar or worse.

Barclay wanted to make sure that UK investigators drew on the firmest science, whether or not it supported the prosecution's case. "What we needed was a sort of unifying group who would work to strict forensic and scientific practices, and I could then use them to coordinate the work of other people or to get them to peer review other people's work," he says. To develop that expertise, Barclay approached Dawson's institute (then called the Macaulay Land Use Research Institute) in 2003 because it was already doing soil analysis for government agencies. That led him to Dawson, who was eager to participate and was already well versed in analyses that could be adapted to forensics work, such as X-ray diffraction, scanning electron microscopy and Fourier transform infrared spectroscopy.

Although the World's End murders had affected her deeply, Dawson had never thought of crime solving as a career. Dirt, on the other hand, came naturally. She grew up on a farm in Angus county, south of Aberdeen. A favourite time of year for her had

been the "tattie holidays", when kids could earn pocket money harvesting tatties, or potatoes. "I just loved that, working outside. You used to make enough money that you could buy a new bike," she says.

After completing her undergraduate degree at Edinburgh, she went on to do a PhD in soil science at the University of Aberdeen, and then worked her way up at the Macaulay, doing a mix of projects related to agriculture and environmental science. When she got the call from Barclay, the offer to work on forensics intrigued her. "It's just another, different sphere of life that the soil can contribute to, really," she says.

Dawson and Barclay put together a team of scientists, investigators and lawyers, which got a grant from the UK Engineering and Physical Sciences Research Council to develop standards for using soil-science techniques in forensic investigations. It was important to get it right, Barclay says, because the high costs of lab work meant that soil science would typically be used only in the most serious cases.

Dawson saw opportunities to develop new techniques, especially for soil's organic matter — the part made up of dead and decomposed plants and organisms. The advantage of studying organic characteristics is that they vary on the scale of centimetres or metres, whereas inorganic components may be broadly the same over kilometres. "The organic takes you to a much finer spatial scale of resolution," she says. When combined with soil-survey databases that document a variety of soil characteristics, that resolution could help investigators to use soil attached to a suspect's shoe or tyre to locate a burial site, for example.

Soon after starting her work in soil forensics, Dawson had a chance to help on a familiar case — the World's End murders. Barclay had been looking into the cold case and he asked Dawson to analyse some dirt and plant material that had come from the bare feet of one of the murdered girls, Helen Scott. But Dawson could not learn much from the soil when she looked at it in 2003 because there was too little to be analysed by all but the most powerful microscopes.

Frustrated but determined to do better, Dawson devoted herself to adapting chromatography and mass spectrometry — techniques she knew from her work in agriculture and environmental science — for use in forensic cases. She was able to substantially reduce the necessary sample size — from about a teaspoon down to about 20 milligrams, roughly the equivalent of a grain of rice. Then she put it into practice. In one case, she was able to use organic characteristics to match soil from getaway vehicles to a crime scene on a remote farm track. When the suspects learned of the soil evidence against them, they pleaded guilty.

Dawson also spread the word about soil forensics by organizing conferences and training in Australia, the United Kingdom, the United States, Russia and elsewhere and by

NATURE.COM
To watch a video on soil forensics, visit.
go.nature.com/o9yulh

holding events for public audiences at home. She collaborated with crime writers such as Ann Cleeves and Stuart MacBride, as well as the BBC, as a way to show the public — and potential jurors — that real forensic evidence is not as clear cut as it is often portrayed on television shows such as *Silent Witness* and *CSI*. At the same time, she demonstrated how useful soil could be in solving crimes. When MacBride fashioned a minor character after her, he gave her a fitting catchphrase: "The soil never lies."

BURIED DRUGS

In her own research, Dawson has kept abreast of developments in soil science, hoping to adapt them to forensics. That is what brought her, clad in hiking boots and gaiters, to a farm southwest of Aberdeen last May. Investigators wanted to know whether a stash of illegal drugs buried at the top of a hill could be linked to dirt samples taken from a suspects' boots and spade.

Dawson hoped to use microbial DNA to solve the case. Soils host vast communities of microbes (up to a couple of billion cells per gram), and those communities can vary on scales as small as millimetres. For several years, forensic scientists have argued that the DNA of those microbial communities could serve as another kind of soil fingerprint (L. M. Macdonald *et al. J. Appl. Microbiol.* **105,** 813–821; 2008).

Such soil fingerprinting has been tried before with mixed success by investigators in Italy, Spain and the Netherlands, which have legal systems that readily consider new forensic techniques. It is much more difficult to use new techniques in trials in common-law countries such as the United Kingdom, United States and Australia, where expert testimony must pass a test to establish the

> ## "THERE'S NO COST THAT YOU SHOULD STOP AT TO TRY AND FIND JUSTICE."

reliability of the methods used. Among other things, judges look for whether the methods have gone through peer review, are generally accepted in the scientific community and have appropriate standards in place.

Dawson, working with a European Union-funded international collaboration called MiSAFE, is trying to meet this high standard for the microbial genetic-profiling technique. That is why the May farm investigation was not a real case, but a mock crime scene. The 'drugs'? White powder. The 'suspects'? Dawson herself and a Hutton staff member.

Back at the Hutton institute, molecular microbiologist Thomas Freitag did the analysis for the mock crime scene: he amplified then chopped up a marker gene that codes for the 16S ribosomal RNA molecule from all the microbes in a clump of soil. Cataloguing these

Police search fields in 1977 near Haddington, Scotland, for the body of murdered teenager Helen Scott.

fragments by length leads to a "coarse portrait" of the community, Freitag says. No two clumps' profiles will match exactly, but clumps that are close together should be roughly similar.

Because the profiling method uses the same technology that forensic units are using today to analyse human DNA, forensic scientists worldwide should be able easily adopt it, says MiSAFE coordinator Edouard Jurkevitch of the Hebrew University of Jerusalem.

In the case of the mock crime scene, the tests performed as hoped. Freitag's analysis revealed, correctly, that the spade belonging to the suspect had been used at the drug-burial site. Dawson presented early results from MiSAFE at a meeting on forensic geosciences in London last December.

The collaboration is now fine-tuning the approach. But Dawson says that she is confident enough in the technique to add microbial DNA profiling to her arsenal as soon as the project wraps up in May.

Dawson's forensic work eventually led her back to the World's End case. Last year prosecutors brought a suspect, Angus Sinclair, to trial. By then, Sinclair was considered the worst serial killer in Scotland's history. DNA from his semen had previously linked him to the World's End murders but it was not enough to convict him during a trial in 2007.

Prosecutor Deborah Demick called on Dawson to reanalyse the soil and plant material from Scott's bare feet. Sinclair claimed that he and his brother-in-law, Gordon Hamilton (who has since died), had had consensual sex with the two girls in their van that night, which explained the presence of his semen on Scott's coat. But Sinclair said that the girls were "alive and unharmed" when Hamilton had dropped him off at a fishing spot, and that Hamilton must have killed the girls later on his own.

By 2014, forensic science had finally advanced enough to refute this alibi and construct an alternative timeline. Investigators reanalysed DNA from inside the knots binding the girls' limbs, showing that Sinclair had helped to tie them. They also concluded that

Sinclair had deposited his semen on Scott's coat just minutes before he left her body in the field.

The debris pressed into Scott's soles helped to fill in the details — and the brutality — of those final minutes, Dawson told the jury on 22 October. Some bits of soil contained traces of plant wax that matched the wheat field where Scott's body was found; other bits matched its grassy border. A similar conclusion emerged from studies of the husks and grains recovered from the dirt on her feet. "The pattern of the soil on her feet suggested that she had walked or stood in that particular field," Dawson says.

The prosecutors used this and other evidence to argue that Sinclair had helped to tie up Scott and forced her to walk from the van into the field, where he beat and strangled her to death. On 14 November, just over 37 years after the attack, the jury convicted Sinclair of the double murder. His life sentence of 37 years was Scotland's longest ever, and the media hailed the historic case as a triumph of forensic science. Dawson's contribution was small but crucial, Demick says, because it "enabled the Crown to have a clear narrative of events and emphasize to the jury the sheer horror of what had happened to Miss Scott — being walked into the field to her death".

For Dawson, the case was poignant and powerful. Her career had come full circle and she had helped to ensure that Sinclair would never again terrify young women. Dawson thought of her own daughters, and how one had almost died several years earlier of leukaemia. She thought of Helen Scott. And she thought of herself at university, studying geology in her dorm room. "Life's so precious, and if it's taken away for whatever reason, particularly if it's taken away by someone else's actions, then I think that there's no cost that you should stop at to try and find justice, to find that person who's done that." ∎

Chelsea Wald *is a freelance reporter in Vienna. She travelled to Scotland with support from a European Geosciences Union fellowship.*

DEGREES OF SEPARATION

Chemists hope to break China's monopoly on rare–earth elements by finding cheap, efficient ways to extract them from ore.

BY XIAOZHI LIM

In July 2010, the Chinese government sent a chill through the world's high-technology industries when it announced a 37% cut in export quotas for rare-earth elements — a group of 17 metallic elements that are essential ingredients in display screens, low-energy lighting, high-powered lasers and a host of other twenty-first-century products.

China has a near monopoly on production of these elements, generating 97% of the world's supply in 2010. So although Beijing said that it was just trying to clean up a particularly dirty sector of its mining industry, the cutback sent rare-earth prices soaring and raised the spectre of major economic disruptions. The reality turned out to be less dire: several Western mining companies have now started producing rare earths, and China, responding to demands from the World Trade Organization, has pledged to end the cutbacks by 2 May this year. Nevertheless, the incident prompted the United States and Europe to launch major research initiatives aimed at securing non-Chinese sources of rare earths. Those programmes are now beginning to achieve results.

Central to the effort is a challenge for chemists. The rare earths are chemically almost identical, generally found together in ore deposits and extremely difficult to separate: the standard method involves some 300 steps and the copious use of hazardous chemicals. China has a network of extraction plants that can undercut any other producer in the world, thanks to the country's historically lax attitude towards costly environmental safeguards, along with factories that incorporate the elements into devices. But if chemists can come up with easier, faster, greener and, above all, cheaper extraction methods, then the balance could shift. Other countries could afford to exploit their own rare-earth deposits, and to recover rare-earth elements from their electronic waste.

"For any Western company ever to be competitive," says Jack Lifton, a consultant for the rare-earth mining industry who works in Farmington Hills, Michigan, "it has to find ways to produce the individual rare earths at lower cost than China."

A worker pours lanthanum into a mould in China.

MASH-UP OF METALS

Disentangling the rare earths is so hard that it took chemists more than a century to identify and name them all: the first, yttrium, was discovered in 1794, yet the last two, lutetium and ytterbium, were not separated until 1907 (see 'Birds of a feather'). Early manufacturers did not bother purifying the metals. The first applications, early in the twentieth century, used a blended 'mischmetal' in cigarette lighter flints and tracer bullets.

Individual rare-earth elements did not begin to find uses until after the Second World War, when the problem of how to separate them caught the attention of Frank Spedding, a chemist at what is now Iowa State University in Ames who had pioneered uranium purification for the Manhattan Project. Spedding perfected a method called ion-exchange chromatography, in which a mixture of rare earths was washed through a vertical glass column packed with polymer beads. The rare earths stuck to the beads, and were then washed off with dilute citric acid.

Spedding adjusted the acid's pH so that each element would dissolve from the polymer at a slightly different rate; as a result, the ions travelled down the column at different speeds before emerging at the bottom in bands of slightly different composition. With enough repetitions, researchers could separate high-purity elements in sufficient quantities to study their properties.

This led to an explosion in rare-earth applications, starting in the early 1960s with the discovery that a pinch of europium oxide mixed with other materials made a bright red colour for television screens. By 1965, soaring demand had led Western mining companies to expand europium production and construct the first separation plants. Because ion exchange was unsuited to large-scale production, the plants used a method called solvent extraction. A mixture of rare earths is dissolved in water, then shaken with organic solvents containing extractants — compounds that bind to some rare-earths more strongly than others. The rare earths move into the organic solvent, and after the mixture has settled, they can be chemically removed from the extractants and redissolved in water. It takes hundreds of repetitions to isolate the elements.

Over the next few decades, Chinese chemical engineers began to up the solvent-extraction technology for use at their own facilities, which were soon able to sell individual rare earths of higher purity and at lower prices than Western producers could. By 1999, China had near-total control of the global rare-earth supply.

CHEAPER, FASTER, CLEANER

One way to change that situation is to step up production at the few solvent-extraction plants outside China, while finding ways to make the method cheaper and more efficient. "Changing the chemistry is a lot easier than changing the infrastructure," says Alex King, director of the Critical Materials Institute (CMI): a US$120-million research hub headquartered in Ames, Iowa, that was set up by the US Department of Energy in 2013 to address supply problems in rare earths and other materials.

Some researchers are trying to change the chemistry by finding extractants that can do a better job of differentiating between the rare earths. It is not easy, says Scott Herbst, a chemical engineer at the CMI who is doing just that at Idaho National Laboratory outside Idaho Falls. "You're separating almost apples from apples." He and his colleagues are tackling the problem both theoretically, by trying to design better molecules with computational modelling, and empirically, seeing whether they can adapt extractants developed for other industries.

Other researchers are looking for better solvents. At the Catholic University of Leuven (KU Leuven) in Belgium, for example, chemist Koen Binnemans is making and testing a variety of ionic liquids: salts that are molten at room temperature. These typically consist of a large organic molecule that carries an electric charge, coupled with a small inorganic ion of the opposite charge. Binnemans says that such ionic liquids are safer, less volatile and more recyclable than the organic solvents most commonly used in industry — not to mention that they can hold around six times the number of dissolved rare-earth ions. Binnemans is trying to develop ionic solvents that can also function as extractants.

Some companies are adapting separation methods from other industries. "If they can cut the cost, they will be very competitive," says Lifton. At one deposit in Alaska, for example, Ucore Rare Metals of Bedford, Canada, has turned ore into gram-scale quantities of 99%-pure individual rare-earth elements, using molecular recognition technology. Developed by the IBC Advanced Technologies in American Fork, Utah, this technique has been used industrially to remove bismuth impurities from copper, and to recover platinum-group metals from scrap catalytic converters.
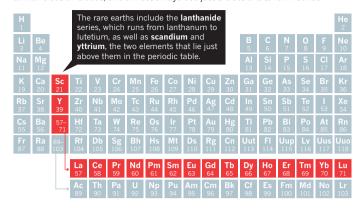
In the Ucore system, a solution of mixed rare earths passes in sequence through 17 different columns, each loaded with a compound tailored to bind to a specific element. That element can then be extracted in 99%-pure form by rinsing the column with dilute acid. Ucore says that the process requires no more than a few repetitions, depending on the desired purity, and so has the potential to be both efficient and environmentally friendly. The company is working with IBC scientists to prove the technology at a pilot plant over the next few months.

BEGIN AGAIN

Digging for rare earths is not always an option — particularly in Europe, where deposits are few and opposition to mining is widespread. But developed nations all over the world have large potential sources of rare earths in the form of used electrical devices. Most

BIRDS OF A FEATHER

The 17 rare-earth elements are chemically similar because they are surrounded by similar electron clouds, and almost always use just 3 electrons to form bonds.

The rare earths include the **lanthanide** series, which runs from lanthanum to lutetium, as well as **scandium** and **yttrium**, the two elements that lie just above them in the periodic table.

fluorescent lights contain europium, yttrium and terbium, for example, and strong permanent magnets typically contain neodymium and dysprosium. "When these reach end of life, we still send them to China" for recycling, says Tom Van Gerven, a chemical engineer at KU Leuven. "We would like to stop, and we can do that if we perform the recycling in Europe."

In recycling, the challenge lies in extracting dilute resources: electronic waste typically has a lower concentration of rare earths than ore does. But any one component also tends to have fewer elements to separate.

Usually, magnets must be dissolved in strong acids to extract the rare earths. Van Gerven is looking for an alternative method: he blasts the solid magnets with ultrasound, commonly used in chemistry laboratories to clean apparatus by dislodging particles from its surfaces. Although experiments are still ongoing, he hopes that the ultrasound will erode the magnet surfaces, allowing extractants to pull out the rare-earth elements without the need to dissolve the entire object.

The CMI also has a team focused on recycling. In one project, extractants are bound to a membrane, where they can catch the rare earths as a solution flows past. This technique can separate elements from even very dilute solutions, says project leader Eric Peterson, who is based at Idaho National Laboratory. He expects full commercialization in the next year or two.

The real-world effect of all of this research is difficult to predict — not least because China is a moving target. Beijing is working hard to reform and strengthen control over its rare-earth industry by consolidating 140 rare-earth companies into 6, cracking down on illegal mining, eliminating unneeded separation plants to reduce supply and raise prices, and conducting environmental remediation to make the industry profitable and socially acceptable. "It was a cowboy industry," says Lifton. "They're reining it in."

These actions could add to the operating cost of China's separation industry, and make it easier for Western plants to compete. But separation is not the whole story, says King. Western manufacturers still have nothing to rival China's network of factories that turn separated rare earths into components such as monitors, magnets and lights. "All of the links in the chain have to be there," he says. Until such plants are built in the West, rare-earth producers such as Molycorp of Greenwood Village, Colorado, and Lynas of Perth, Australia, will have few places to sell their output — except China. And with global demand threatening to outstrip that country's mining capacity, Chinese producers are already looking to feed their factories by mining deposits of rare-earth elements overseas.

So even if Western producers manage to develop separation processes that can produce the individual elements cheaply, economics might eventually force an ironic turnaround — with China importing rare earths from the West. ■

XiaoZhi Lim *is a freelance writer in Singapore.*

COMMENT

The Leiden Manifesto for research metrics

Use these ten principles to guide research evaluation, urge **Diana Hicks**, **Paul Wouters** and colleagues.

Data are increasingly used to govern science. Research evaluations that were once bespoke and performed by peers are now routine and reliant on metrics[1]. The problem is that evaluation is now led by the data rather than by judgement. Metrics have proliferated: usually well intentioned, not always well informed, often ill applied. We risk damaging the system with the very tools designed to improve it, as evaluation is increasingly implemented by organizations without knowledge of, or advice on, good practice and interpretation.

Before 2000, there was the Science Citation Index on CD-ROM from the Institute for Scientific Information (ISI), used by experts for specialist analyses. In 2002, Thomson Reuters launched an integrated web platform, making the Web of Science database widely accessible. Competing citation indices were created: Elsevier's Scopus (released in 2004) and Google Scholar (beta version released in 2004). Web-based tools to easily compare institutional research productivity and impact were introduced, such as InCites (using the Web of Science) and SciVal (using Scopus), as well as software to analyse individual citation profiles using Google Scholar (Publish or Perish, released in 2007).

In 2005, Jorge Hirsch, a physicist at the University of California, San Diego, proposed the h-index, popularizing citation counting for individual researchers. Interest in the journal impact factor grew steadily after 1995 (see 'Impact-factor obsession').

Lately, metrics related to social usage ▶

▶ and online comment have gained momentum — F1000Prime was established in 2002, Mendeley in 2008, and Altmetric.com (supported by Macmillan Science and Education, which owns Nature Publishing Group) in 2011.

As scientometricians, social scientists and research administrators, we have watched with increasing alarm the pervasive misapplication of indicators to the evaluation of scientific performance. The following are just a few of numerous examples. Across the world, universities have become obsessed with their position in global rankings (such as the Shanghai Ranking and *Times Higher Education*'s list), even when such lists are based on what are, in our view, inaccurate data and arbitrary indicators.

Some recruiters request *h*-index values for candidates. Several universities base promotion decisions on threshold *h*-index values and on the number of articles in 'high-impact' journals. Researchers' CVs have become opportunities to boast about these scores, notably in biomedicine. Everywhere, supervisors ask PhD students to publish in high-impact journals and acquire external funding before they are ready.

In Scandinavia and China, some universities allocate research funding or bonuses on the basis of a number: for example, by calculating individual impact scores to allocate 'performance resources' or by giving researchers a bonus for a publication in a journal with an impact factor higher than 15 (ref. 2).

In many cases, researchers and evaluators still exert balanced judgement. Yet the abuse of research metrics has become too widespread to ignore.

We therefore present the Leiden Manifesto, named after the conference at which it crystallized (see http://sti2014.cwts.nl). Its ten principles are not news to scientometricians, although none of us would be able to recite them in their entirety because codification has been lacking until now. Luminaries in the field, such as Eugene Garfield (founder of the ISI), are on record stating some of these principles[3,4]. But they are not in the room when evaluators report back to university administrators who are not expert in the relevant methodology. Scientists searching for literature with which to contest an evaluation find the material scattered in what are, to them, obscure journals to which they lack access.

We offer this distillation of best practice in metrics-based research assessment so that researchers can hold evaluators to account, and evaluators can hold their indicators to account.

TEN PRINCIPLES

1 **Quantitative evaluation should support qualitative, expert assessment.** Quantitative metrics can challenge bias tendencies in peer review and facilitate deliberation. This should strengthen peer review, because making judgements about colleagues is difficult without a range of relevant information. However, assessors must not be tempted to cede decision-making to the numbers. Indicators must not substitute for informed judgement. Everyone retains responsibility for their assessments.

2 **Measure performance against the research missions of the institution, group or researcher.** Programme goals should be stated at the start, and the indicators used to evaluate performance should relate clearly to those goals. The choice of indicators, and the ways in which they are used, should take into account the wider socio-economic and cultural contexts. Scientists have diverse research missions. Research that advances the frontiers of academic knowledge differs from research that is focused on delivering solutions to societal problems. Review may be based on merits relevant to policy, industry or the public rather than on academic ideas of excellence. No single evaluation model applies to all contexts.

> *"Simplicity is a virtue in an indicator because it enhances transparency."*

3 **Protect excellence in locally relevant research.** In many parts of the world, research excellence is equated with English-language publication. Spanish law, for example, states the desirability of Spanish scholars publishing in high-impact journals. The impact factor is calculated for journals indexed in the US-based and still mostly English-language Web of Science. These biases are particularly problematic in the social sciences and humanities, in which research is more regionally and nationally engaged. Many other fields have a national or regional dimension — for instance, HIV epidemiology in sub-Saharan Africa.

This pluralism and societal relevance tends to be suppressed to create papers of interest to the gatekeepers of high impact: English-language journals. The Spanish sociologists that are highly cited in the Web of Science have worked on abstract models or study US data. Lost is the specificity of sociologists in high-impact Spanish-language papers: topics such as local labour law, family health care for the elderly or immigrant employment[5]. Metrics built on high-quality non-English literature would serve to identify and reward excellence in locally relevant research.

4 **Keep data collection and analytical processes open, transparent and simple.** The construction of the databases required for evaluation should follow clearly stated rules, set before the research has been completed. This was common practice among the academic and commercial groups that built bibliometric evaluation methodology over several decades. Those groups referenced protocols published in the peer-reviewed literature. This transparency enabled scrutiny. For example, in 2010, public debate on the technical properties of an important indicator used by one of our groups (the Centre for Science and Technology Studies at Leiden University in the Netherlands) led to a revision in the calculation of this indicator[6]. Recent commercial entrants should be held to the same standards; no one should accept a black-box evaluation machine.
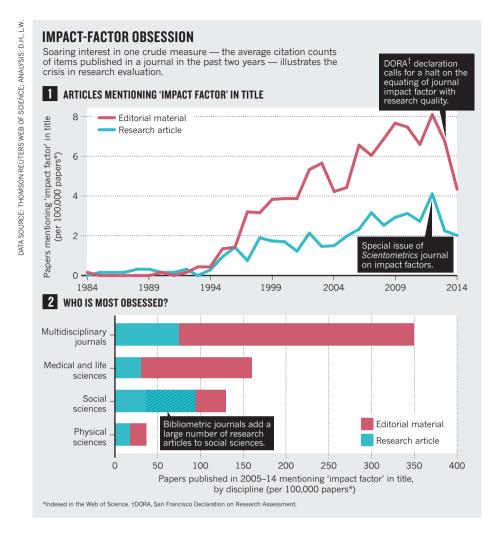
Simplicity is a virtue in an indicator because it enhances transparency. But simplistic metrics can distort the record (see principle 7). Evaluators must strive for balance — simple indicators true to the complexity of the research process.

5 **Allow those evaluated to verify data and analysis.** To ensure data quality, all researchers included in bibliometric studies should be able to check that their outputs have been correctly identified. Everyone directing and managing evaluation processes should assure data accuracy, through self-verification or third-party audit. Universities could implement this in their research information systems and it should be a guiding principle in the selection of providers of these systems. Accurate, high-quality data take time and money to collate and process. Budget for it.

6 **Account for variation by field in publication and citation practices.** Best practice is to select a suite of possible indicators and allow fields to choose among them. A few years ago, a European group of historians received a relatively low rating in a national peer-review assessment because they wrote books rather than articles in journals indexed by the Web of Science. The historians had the misfortune to be part of a psychology department. Historians and social scientists require books and national-language literature to be included in their publication counts; computer scientists require conference papers be counted.

Citation rates vary by field: top-ranked journals in mathematics have impact factors of around 3; top-ranked journals in cell biology have impact factors of about 30. Normalized indicators are required, and the most robust normalization method is based on percentiles: each paper is weighted on the basis of the percentile to which it belongs in the citation distribution of its field (the top 1%, 10% or 20%, for example). A single highly cited publication slightly improves the position of a university in a ranking that

IMPACT-FACTOR OBSESSION

Soaring interest in one crude measure — the average citation counts of items published in a journal in the past two years — illustrates the crisis in research evaluation.

1 ARTICLES MENTIONING 'IMPACT FACTOR' IN TITLE

DORA[†] declaration calls for a halt on the equating of journal impact factor with research quality.

Special issue of *Scientometrics* journal on impact factors.

2 WHO IS MOST OBSESSED?

Bibliometric journals add a large number of research articles to social sciences.

Editorial material
Research article

Papers published in 2005–14 mentioning 'impact factor' in title, by discipline (per 100,000 papers*)

*Indexed in the Web of Science. †DORA, San Francisco Declaration on Research Assessment.

is based on percentile indicators, but may propel the university from the middle to the top of a ranking built on citation averages[7].

7 Base assessment of individual researchers on a qualitative judgement of their portfolio. The older you are, the higher your *h*-index, even in the absence of new papers. The *h*-index varies by field: life scientists top out at 200; physicists at 100 and social scientists at 20–30 (ref. 8). It is database dependent: there are researchers in computer science who have an *h*-index of around 10 in the Web of Science but of 20–30 in Google Scholar[9]. Reading and judging a researcher's work is much more appropriate than relying on one number. Even when comparing large numbers of researchers, an approach that considers more information about an individual's expertise, experience, activities and influence is best.

8 Avoid misplaced concreteness and false precision. Science and technology indicators are prone to conceptual ambiguity and uncertainty and require strong assumptions that are not universally accepted. The meaning of citation counts, for example, has long been debated. Thus,

best practice uses multiple indicators to provide a more robust and pluralistic picture. If uncertainty and error can be quantified, for instance using error bars, this information should accompany published indicator values. If this is not possible, indicator producers should at least avoid false precision. For example, the journal impact factor is published to three decimal places to avoid ties. However, given the conceptual ambiguity and random variability of citation counts, it makes no sense to distinguish between journals on the basis of very small impact factor differences. Avoid false precision: only one decimal is warranted.

9 Recognize the systemic effects of assessment and indicators. Indicators change the system through the incentives they establish. These effects should be anticipated. This means that a suite of indicators is always preferable — a single one will invite gaming and goal displacement (in which the measurement becomes the goal). For example, in the 1990s, Australia funded university research using a formula based largely on the number of papers published by an institute. Universities could calculate the 'value' of a paper in

a refereed journal; in 2000, it was Aus$800 (around US$480 in 2000) in research funding. Predictably, the number of papers published by Australian researchers went up, but they were in less-cited journals, suggesting that article quality fell[10].

10 Scrutinize indicators regularly and update them. Research missions and the goals of assessment shift and the research system itself co-evolves. Once-useful metrics become inadequate; new ones emerge. Indicator systems have to be reviewed and perhaps modified. Realizing the effects of its simplistic formula, Australia in 2010 introduced its more complex Excellence in Research for Australia initiative, which emphasizes quality.

NEXT STEPS

Abiding by these ten principles, research evaluation can play an important part in the development of science and its interactions with society. Research metrics can provide crucial information that would be difficult to gather or understand by means of individual expertise. But this quantitative information must not be allowed to morph from an instrument into the goal.

The best decisions are taken by combining robust statistics with sensitivity to the aim and nature of the research that is evaluated. Both quantitative and qualitative evidence are needed; each is objective in its own way. Decision-making about science must be based on high-quality processes that are informed by the highest quality data. ∎

Diana Hicks *is professor of public policy at the Georgia Institute of Technology, Atlanta, Georgia, USA.* **Paul Wouters** *is professor of scientometrics and director,* **Ludo Waltman** *is a researcher, and* **Sarah de Rijcke** *is assistant professor, at the Centre for Science and Technology Studies, Leiden University, the Netherlands.* **Ismael Rafols** *is a science-policy researcher at the Spanish National Research Council and the Polytechnic University of Valencia, Spain.*
e-mail: diana.hicks@pubpolicy.gatech.edu

1. Wouters, P. in *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).
2. Shao, J. & Shen, H. *Learned Publ.* **24,** 95–97 (2011).
3. Seglen, P. O. *Br. Med. J.* **314,** 498–502 (1997).
4. Garfield, E. *J. Am. Med. Assoc.* **295,** 90–93 (2006).
5. López Piñeiro, C. & Hicks, D. *Res. Eval.* **24,** 78–89 (2015).
6. van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. *J. Informetrics* **4,** 431–435 (2010).
7. Waltman, L. *et al. J. Am. Soc. Inf. Sci. Technol.* **63,** 2419–2432 (2012).
8. Hirsch, J. E. *Proc. Natl Acad. Sci. USA* **102,** 16569–16572 (2005).
9. Bar-Ilan, J. *Scientometrics* **74,** 257–271 (2008).
10. Butler, L. *Res. Policy* **32,** 143–155 (2003).

Global monitoring of atmospheric nitrogen dioxide reveals pollution hotspots.

Five priorities for the UN Sustainable Development Goals

Restructure data–gathering and evaluation networks to address climate change, energy, food, health and water provision, say **Yonglong Lu** and colleagues.

This week, the United Nations is deliberating in New York how to implement the 17 Sustainable Development Goals (SDGs) that it will launch formally in September. Science must be at the heart of its plans.

The SDGs place greater demands on the scientific community than did the Millennium Development Goals (MDGs), which they replace[1]. Addressing climate change, renewable energy, food, health and water provision requires coordinated global monitoring and modelling of many factors — social, economic and environmental.

Much remains to be done: the 17 goals comprise 169 targets, 91 of which need to be specified in more detail. Metrics need to be developed to measure progress towards the targets on local, national, regional and global levels and across sectors. Monitoring and evaluation procedures and standards need to be set up.

To guide action, the connections between targets need to be better understood. Some synergies and trade-offs depend on scale — for instance, greater fertilizer use might increase food production and incomes locally, but would exacerbate pollution. Climate-change mitigation tends to happen at a local scale, but the consequences are global.

We lay out five priorities for how the scientific community should participate in this process, based on the findings of a scientific review[1] of the draft SDGs conducted by the International Council for Science (ICSU).

FIVE PRIORITIES

Devise metrics. Scientists, social scientists and economists need to design a set of practical indices for tracking progress on each SDG. Ensuring access to sustainable and modern energy for all (goal 7), for example, will require indicators of improvements in energy efficiency and carbon savings from renewable-energy technologies (see go.nature.com/pkij7y). Parameters other than just economic growth must be included, such as income inequality, carbon emissions, population and lifespans[1].

Existing methodologies can form the starting points, including environmental impact assessment, natural-asset valuation, cost–benefit analysis and life-cycle costing. Ambiguous terms in the wording of the goals, such as 'sustainable', 'efficient' and 'substantial', need to be defined quantitatively so that the goals can be measurable, comparable and achievable. Scientific analyses of the effectiveness of different scenarios should inform the metrics.

Establish monitoring mechanisms. Governments and researchers must decide which values need to be tracked, and set up systems to acquire the data. Quantities such as water and energy consumption, emissions and health impacts need to be monitored, as do scientific variables such as water pH, turbidity and metal concentrations. Social scientists should propose what sorts of data on behaviour, values and beliefs should be collected and analysed, how and by whom[1].

Analysis and interpretation must be provided at the same time, ideally by an independent government-backed organization, to consider the data in context. In water-quality monitoring, for example, measurements of physical, chemical and biological features such as pH and chemical oxygen demand are compared against national or regional water-quality standards that relate to impacts on human and ecosystem health[2].

Global collaboration between governments and scientific bodies will be essential in setting up monitoring programmes and in assisting developing nations to implement them[3].

Evaluate progress. Scientists should help to choose criteria — such as the water-quality standards — against which progress towards the goals is judged, based on accepted principles of good practice or governance such as social equality or cost-effectiveness.

A peer-review mechanism should be established through the UN platform for intergovernmental negotiations to evaluate the performance and implementation of SDG projects and policies every 3–5 years — and suggest reforms where necessary.

The Intergovernmental Panel on Climate Change, the Intergovernmental Platform on Biodiversity and Ecosystem Services and other scientific-assessment bodies should set up task forces to decide how they can evaluate relevant aspects of the goals, either by extending their remits or by sharing their experiences with a central SDG assessment body[3].

The SDG evaluators must also decide how to incorporate the contributions of regions, cities, companies and others into national and international pledges; consider national or local circumstances when evaluating progress; and check whether sustainable development has been incorporated into planning processes and strategies at all levels.

Enhance infrastructure. Earth observation, ground-based monitoring and information-processing capabilities need to be expanded to give better global coverage, to allow direct comparisons of data by using similar instruments, and to store, analyse and share data.

Much of this can be achieved if the SDG process engages with Future Earth, a ten-year initiative for global-sustainability research launched at the Rio+20 UN conference in 2012. Future Earth's observing networks, high-performance computing, Earth-system models, theoretical frameworks, data-management systems and research infrastructures need to be enhanced to track human dimensions and societal changes. The Science and Technology Alliance for Global Sustainability should help Future Earth to revise its agenda to include observations of production, consumption and urbanization.

The ICSU should work with international bodies such as the World Meteorological Organization, the UN Educational, Scientific and Cultural Organization, and the UN Environment Programme to shape global monitoring networks to serve the SDGs. This will include changing the missions, objectives and targets of existing global observation systems, including those for climate, oceans and ecosystems.

Space-based monitoring coupled with on-the-ground observation is crucial for tracking changes on large scales — in natural-resource availability, landscape patterns and management, and social structure. Government investments need to be maintained or increased.

Data gathering using smart phones and miniature sensors can capture point-source industrial emissions, waste dumping and unsustainable consumption. Citizen science provides new opportunities. For example, starting in 2008, Beijing residents' use of Twitter to share air-pollution data obtained from home-based monitoring equipment pressured the municipality to improve its air-quality monitoring. Such actions also serve as a check on reporting bias from governments, companies and others.

Geographical information systems will be needed to host and share data from the observing networks. Image processing, simulation and decision-making tools are also needed to support sustainability planning, management and enforcement. These should be openly accessible so that citizens can make informed decisions about, say, how to improve air quality or ocean health[4]. Businesses can track their environmental impacts and project future supplies[5].

Developing countries will need to collaborate with developed countries to build capacity in observing, data mining and statistics[6]. Bodies such as the G20 Development Working Group and the InterAcademy Panel should help with this.

Standardize and verify data. Countries' capabilities to acquire and process economic and social data vary greatly. The likelihood of collecting wrong or useless information is high, owing to a lack of standards and consistent methods and instrumentation. Scientists and governments need to design monitoring and sampling approaches with robustness in mind, and to verify data.

Cooperation among authorities and agencies is more effective if partners collect data according to agreed definitions, specifications, methodologies and formats. The Partnership in Statistics for Development in the 21st Century (PARIS21) aims to strengthen national statistical systems to monitor global goals, by promoting such shared principles[3].

Checks must be made, for instance, by implementing complementary monitoring systems, comparing results from airborne and land-based surveillance or by inspecting data-collection methods. Shared online reporting platforms — perhaps modelled on the Google Earth Engine or the NASA Earth Exchange — should provide quality-control services for data from different sources[7,8].

All SDG data must be open access and released as soon as possible. Governments, scientists and corporations will need to commit to long-term funding for the collection, calibration and release of data sets, which should be discussed in the SDG political-negotiation process. These commitments could then be checked at the high-level political forum held every year. Sanctions such as economic penalties will need to be considered for governments and companies that do not comply.

Scientists need to support the SDGs. We must help to integrate monitoring and evaluation mechanisms into policy-making at all levels and ensure that information about our planet is easily available to all. ∎

> *"Developing countries will need to collaborate with developed countries to build capacity."*

Yonglong Lu *is professor at the State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China.* **Nebojsa Nakicenovic** *is professor and former deputy director at the International Institute for Applied Systems Analysis (IIASA), Vienna, Austria.* **Martin Visbeck** *is professor at the GEOMAR Helmholtz Centre for Ocean Research and Kiel University, Germany.* **Anne-Sophie Stevance** *is a science officer at the International Council for Science, Paris, France.*
e-mail: yllu@rcees.ac.cn

1. International Council for Science, International Social Science Council. *Review of Targets for the Sustainable Development Goals: The Science Perspective* (ICSU, 2015).
2. *Australian and New Zealand Guidelines for Fresh and Marine Water Quality* (Australian and New Zealand Environment and Conservation Council, Agriculture and Resource Management Council of Australia and New Zealand, 2000).
3. Organisation for Economic Co-operation and Development. *Strengthening National Statistical Systems to Monitor Global Goals* (OECD, 2013).
4. German Advisory Council on Global Change. *World in Transition: Governing the Marine Heritage* (WBGU, 2013).
5. European Environmental Agency. *Shared Environmental Information System (SEIS)* (EEA, 2011).
6. *Saint Petersburg Development Outlook*: Annex to the G20 Leaders' Declaration (G20, 2013).
7. United Nations Environment Programme. *Embedding the Environment in Sustainable Development Goals* (UNEP, 2013).
8. UN Independent Expert Advisory Group on the Data Revolution for Sustainable Development. *A World That Counts: Mobilising the Data Revolution for Sustainable Development* (UN IEAG, 2014).

The ecosystem services — and value — provided by mangrove forests range from coastal protection from storms to natural fisheries.
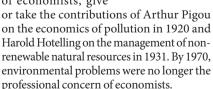
ENVIRONMENTAL ECONOMICS

Pricing the planet

Nick Hanley weighs up a study that probes the economic value of nature.

Can economics help to save us from environmental catastrophe? In *Natural Capital*, economist Dieter Helm offers a timely reminder of the contribution that his discipline can make to understanding and solving environmental problems. The book hinges on the economic value of gifts of nature, from oil fields to wetlands, which in combination with inputs such as labour and produced capital provide humanity with valuable benefits from ecosystem services. Helm's main message is that the apparent conflict between economic growth and environmental quality can be managed by preventing declines in this natural capital.

The concept of natural capital has quite a backstory, although Helm does not delve into it too deeply. The environment was of central concern to the classical economists of the nineteenth century. The scarcity of productive agricultural land and coal reserves, for instance, was seen as a brake on economic growth by pioneers such as David Ricardo and John Stuart Mill. Slightly earlier, Thomas Robert Malthus had famously predicted a gloomy future as a result of the conflict between an exponentially growing human population and the fixed amount of farmland, which meant that food supplies could not keep up. However, disaster failed to materialize, population and average living standards continued to rise, and the environment largely disappeared from the thinking and writing of economists, give or take the contributions of Arthur Pigou on the economics of pollution in 1920 and Harold Hotelling on the management of non-renewable natural resources in 1931. By 1970, environmental problems were no longer the professional concern of economists.

Then everything changed. In 1972, environmental scientist Donella Meadows and co-authors published *The Limits to Growth* (Universe), commissioned by the Club of Rome think tank. This influential book used systems dynamics modelling to predict probable future paths for global population, food production and pollution. Some of these

Natural Capital: Valuing the Planet
DIETER HELM
Yale Univ. Press: 2015.

predictions recalled Malthus. This, along with the two oil-price peaks of the 1970s and growing public interest in the damaging implications of economic growth, returned the environment to centre stage in economics. *The Journal of Environmental Economics and Management* was founded in 1974.

The late, great David Pearce was perhaps the first academic economist to convince people outside the field of its relevance for understanding the relationships between people, money and the environment — and for developing tools to help to manage the apparent conflict between economic growth and environmental quality. The influence of his 1989 *Blueprint for a Green Economy* (Routledge), co-written with Anil Markandya and Edward Barbier, reached beyond academia and government to the informed public. Pearce had three main messages. First, economic benefits from the environment need to be measured and recognized. Second, economics could improve environmental policy by developing market-like mechanisms through which a price could be put on pollution. Third, national accounting conventions needed to show up the gains and losses in a country's natural capital over time.

These ideas are all taken up and expanded in *Natural Capital*. This is important, because the empirical evidence is that most countries do not account for the economic value of depreciating natural capital; nor have they put in place measures to hold the line. Helm's arguments bring the main problem raised by Malthus into a sharp new focus. Given current rates of world economic growth, incredible numbers of people, demands for resources and levels of pollution now loom, increasing pressures on ecosystems and biodiversity. Evidence is growing of the importance of ecosystem services such as clean water and pollination, and of the erosion of human well-being that results when those services are disturbed. That does not mean that economic growth should be stopped (even if that were possible), but it does demand a fundamental change in government policies globally.

As Helm drives home, these changes relate most fundamentally to a new goal of economic policy: keeping natural capital from declining. Many of the assets that make up natural capital deliver benefits that the market does not value, but which are important for well-being. So adopting such a policy would mean that as a country depletes its oil reserves, for example, it would reinvest a proportion of the returns from this activity in promoting renewable alternatives.

That demands a number of moves. A country must change the way it undertakes its national accounting to reflect the year-on-year changes in the value of all of its assets, including natural capital; it must tax pollution while removing perverse subsidies for activities that deplete natural capital; it must enforce strict limits on the use of renewable resources to maintain them above critical thresholds; it must require general offsetting of the negative effects of infrastructure projects. Moreover, it must increase the provision of public goods such as national parks and green spaces.

These are not new ideas (most were discussed in *Blueprint for a Green Economy*), but *Natural Capital* provides a very useful update and pulls together the past 20 years of economic insight in language that non-economists will easily understand. For example, since 1989 economists have made great progress in estimating the values of ecosystem-service benefits. Helm has thought carefully about the practicalities of tracking changes in natural capital, of funding reinvestment in habitats, and of prioritizing actions through a focus on thresholds. As such, the book is a valuable contribution, written by an author who knows his subject and cares deeply about his message. ∎

Nick Hanley *is a professor of environmental economics at the University of St Andrews, UK. e-mail: ndh3@st-andrews.ac.uk*

LIBRARY SCIENCE

The word on our archival future

Michael Lesk assesses a work on the fate of the library at a time of economic and technological upheaval.

John Palfrey loves libraries and is optimistic about their future. In *BiblioTech*, he describes that future. In his vision, public libraries, as a mainstay of informed democratic societies, will share access to online material as well as to the physical books and spaces that resonate emotionally with users.

Palfrey — who helped to set up the Digital Public Library of America at Harvard University in Boston, Massachusetts — shows that most libraries are facing a "perfect storm": waning use, inadequate resources and users satisfied with easy personal access to online information. Although the book does not focus on academic libraries, US research libraries saw a 69% decline in the number of queries handled between 1991 and 2012. US research librarians answer some 7 million questions each year; Internet search engines handle that many every two minutes. Academic-library budgets are shrinking: spending has dropped from 3.7% of the total budget of a typical US university in the early 1980s to 1.8% in 2011.

Roughly half of an academic-library budget is spent on acquisitions, and that is increasingly consumed by journal-subscription prices. However, open-access publishers such as the Public Library of Science (PLOS) provide information with no reader fee, changing the economics entirely.

As Palfrey notes, school libraries — which still have a key role in teaching scholarly habits and propelling children towards a life in science — are suffering, often losing their staff. The 2010 Google Map 'A Nation Without School Librarians' is a grim indicator of the trend, showing the districts that have eliminated certified school-librarian positions. As for public libraries, the Pew Research Center reports that only about half of people in the United States used one in 2013. Palfrey is fearful that in the future, not everyone will have access to reliable information, and he is unconvinced that poorer people and nations will

BiblioTech: Why Libraries Matter More Than Ever in the Age of Google
JOHN PALFREY
Basic: 2015.

> *"The school library should be as important as the school sports team."*

have equal access to information online.

The increasing commercialization of information raises further barriers. Palfrey is especially concerned about the copyright status of out-of-print books. Nobody makes money from them, but legal problems such as the dilemma of 'orphan books' (volumes in copyright, but for which the owner is unlocatable) prevent libraries from providing such volumes online. Finally, Palfrey, echoing the concerns of Google vice-president Vint Cerf among others, feels that digital preservation of library holdings is riskier than traditional methods. In my view, this is not entirely accurate: a single paper copy can fade or go up in smoke, whereas multiple digital copies can stay safe and affordable.

At a time of austerity and belt-tightening, will governments recognize the importance of libraries and librarians? The school library should be as important as the school sports team, I feel. Political will is essential, as is innovative energy among librarians. Palfrey hopes that conventional and new library technologies will sit side by side. This is not common in technological change: car-rental companies do not run livery stables. I suspect that a more likely future is that libraries (and museums) will be divided into the 'wholesalers' that have large historic collections, such as the US Library of Congress or the British Library, and the 'retailers' such as university libraries, which serve faculties.

Anyone interested in the future of libraries — and whether there is one at all — will find much to mull over in this book. I hope its effectiveness will match its enthusiasm. ∎

Michael Lesk *is professor of information science at Rutgers University in New Brunswick, New Jersey, and chair of its department of library and information science. e-mail: lesk@rci.rutgers.edu*

Correspondence

Music inspired Newton's rainbow

Isaac Newton was among the great scientists who took inspiration from music (see *Nature* **519,** 262; 2015). In fact, music drove him to add two new colours to the rainbow.

The medieval rainbow had just five colours: red, yellow, green, blue and violet. Newton added two more — orange and indigo — so that the colours would be "divided after the manner of a Musical Chord" (I. Newton in *Opticks* 4th edn, 127 (William Innys, 1730); see also K. McLaren *Color Res. Application* **10,** 225–229; 1985).

On a seemingly unrelated note, Ed Hawkins and colleagues make a plea to scrap rainbow colour scales in scientific graphics (*Nature* **519,** 291; 2015). They warn that these palettes can "introduce false perceptual thresholds in the data".

It was Newton's perception that first introduced new colour thresholds, although these were subjective rather than false. It gives the lie to the old saying that artists see what they believe, but scientists believe what they see.

Len Fisher *University of Bristol, UK.*
len.fisher@bristol.ac.uk

Disputed start dates for Anthropocene

As members of the Anthropocene Working Group, we contend that the proposed new geological epoch should reflect a unique stratigraphic unit that is characterized by unambiguous, widespread and essentially permanent anthropogenic signatures in rock, glacial ice or marine sediments. We therefore find the two dates chosen by Simon Lewis and Mark Maslin to be questionable candidates for the start of the Anthropocene (*Nature* **519,** 171–180; 2015).

For the first date suggested by the authors, the short-lived decline of atmospheric carbon dioxide that reached its minimum in 1610 is not an ideal stratigraphic marker for an epoch-scale boundary. It is one small dip of several in the Holocene epoch, which began about 11,700 years ago, and is not outside the range of natural variability — in contrast to the signature associated with industrialization. Associated indicators of colonization of the Americas, such as the worldwide spread of pollen from maize (corn), lasted for centuries and so do not represent near-synchronous markers.

By the time of the authors' other suggested date of 1964, the 'great acceleration' in human activity was well under way (W. Steffen *et al. Anthropocene Rev.* **2,** 81–98; 2015). Also, the year 1964 is later than the near-synchronous upward inflections of many physical and socio-economic trends and their respective stratigraphic signals, which date to around 1950 (J. Zalasiewicz *et al. Quat. Int.* http://doi.org/zjf; 2015)

We need further comprehensive analyses of the advantages and limitations of different proposed markers before we can arrive at an effective starting date for the Anthropocene.

Jan Zalasiewicz* *University of Leicester, UK.*
jaz1@leicester.ac.uk

*On behalf of 24 correspondents (see go.nature.com/3z9oju for full list).

Act to staunch loss of research data

Never before have scientists had the ability to generate and collect so much data — recent estimates suggest that the global scientific output is doubling roughly every decade (see L. Bornmann and R. Mutz, preprint at http://arxiv.org/abs/1402.4578v3; 2014, and go.nature.com/nzejwh). It is alarming, therefore, that the odds of data being lost are estimated to increase by 17% in every year after publication (T. H. Vines *et al. Curr. Biol.* **24,** 94–97; 2014). And this does not include the 80% or so of research data that are inaccessible or unpublished (B. P. Heidorn *Libr. Trends* **57,** 280–299; 2008).

Information is lost when researchers fail to store, archive or share their data, for example, and as a result of ageing technology or corruption of data-storage devices. A culture of systematic data curation is needed to stem this loss, but it is not yet in place across research fields — even though curation costs a fraction of the funding used to generate the data in the first place. Standardized protocols would ensure that data are shared and properly curated worldwide.

Global networks such as the Confederation of Open Access Repositories can support research institutions in storing their data. National data services are already providing generic support to researchers (see, for example, go.nature.com/uns6zy). Now, different fields need to converge on common formats for data storage and preservation if such measures are to be effective.

Andrew Gonzalez *McGill University, Montreal, Canada.*
Pedro R. Peres-Neto *University of Quebec in Montreal, Canada.*
andrew.gonzalez@mcgill.ca

China needs more monitoring apps

There are more than one billion mobile devices in China, offering huge potential for citizen scientists to contribute to a cleaner and safer environment. The scientific community should rapidly develop mobile apps to collect and monitor environmental and biodiversity data.

In one example of how China's citizen science could take off, journalist Jing Chai showed how to use a mobile app to follow air quality in the local environment as part of her 'Under the Dome' documentary on smog pollution, which was viewed millions of times in just 24 hours (see go.nature.com/2hj7wa).

Better apps would allow China's 8,000 or so environmental non-governmental organizations to tap into the massive volume of data collected by professionals and citizens. Also, collaborations among all three groups could improve and streamline data management and analysis.

Government financial support for citizen science is currently sparse. Greater investment by large agencies such as the National Natural Science Foundation of China would quickly reap rewards.

Jian Zhang *Aarhus University, Denmark.*
Xiaolei Huang *Fujian Agriculture and Forestry University, China.*
jian@bios.au.dk

Illuminating the past

The technique of optical dating was first reported 30 years ago, and has since revolutionized studies of events that occurred during the past 500,000 years. Here, two practitioners of optical dating assess its impact and consider its future.

RICHARD G. ROBERTS & OLAV B. LIAN

Thirty-year anniversaries are traditionally associated with pearls, which are renowned for the lustre produced by the reflection, refraction and diffraction of light. It is fitting, then, that in this International Year of Light and Light-based Technologies, we also celebrate the dawn of the optical dating technique, first reported three decades ago by David Huntley and colleagues in *Nature*[1]. Optical dating was proposed by the authors as a method for determining the time since wind-blown and water-borne mineral grains were last bleached by the Sun's rays before becoming buried, for example in a sedimentary landform. It has since become an essential arrow in the quiver of scientists worldwide, enabling geological, biological and archaeological events to be placed on a timescale extending from the present to half a million years ago or earlier — well beyond the 50,000-year limit of radiocarbon dating, and without the need for subsequent calibration corrections.

Optical dating exploits the physical properties of light-sensitive electron traps in ubiquitous minerals — chiefly, quartz and feldspar — as atomic 'time capsules'. These traps are rapidly emptied when exposed to sunlight, but steadily refill if mineral grains
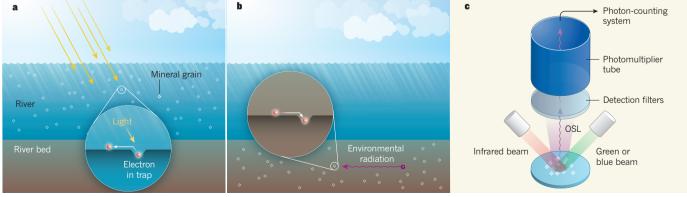
are buried within a deposit and concealed from light, because of the energy received from background environmental radiation (Fig. 1). The time elapsed since the grains were last bleached by sunlight is calculated from a laboratory estimate of the past radiation dose divided by the rate at which ionizing radiation from environmental sources is absorbed by the grains after burial[1–7].

Huntley, together with Ann Wintle, had previously been pivotal in developing reliable procedures for thermoluminescence dating of unheated sediments[8]. This technique is closely related to optical dating, except that thermoluminescence traps are emptied by heating grains — a process that evicts electrons from both optically inert traps and light-sensitive ones. Optical dating, by contrast, enables the latter to be accessed directly. Huntley *et al.*[1] achieved this using the green beam from a powerful argon-ion laser to induce dim, optically stimulated luminescence (OSL) from quartz grains[3]. They then compared this signal with the OSL obtained from grains that had been dosed with radiation in the laboratory, to estimate the past radiation dose and therefore the burial time of the grains.

Their approach was promptly implemented by another team using a similar laser[9], but optical dating spread more widely only after

it was discovered that feldspars are acutely sensitive to infrared stimulation[10], enabling the convenient use of infrared light-emitting diodes (LEDs). By the late 1990s, the technique had matured into a powerful tool for dating sediments from the Quaternary period (the current geological period, which began about 2.6 million years ago), shedding light on the evolution of desert dunes and other landforms and on the timing of past human activities, particularly in Australia and Europe[2].

Applications proliferated after the turn of the millennium, following a decade of development of 'single-aliquot' procedures[2] to determine the burial dose — an idea originally proposed by Huntley and colleagues[1]. The adoption of optical dating by laboratories worldwide was then spurred by several advances: the advent of single-aliquot regenerative-dose (SAR) procedures[11] (which involves making repeated OSL measurements on individual grains or separate groups of grains to obtain many independent estimates of the burial dose for a sediment sample); the use of established statistical methods to analyse OSL data[12]; and the incorporation of sufficiently bright LEDs and compact solid-state lasers to stimulate quartz and feldspar grains in purpose-built, automated instruments[13].

The resulting studies have addressed

Figure 1 | Electron traps as timekeepers in mineral grains. a, Mineral grains are exposed to sunlight when transported by air or water, or when deposited on the ground. Electrons caught in light-sensitive traps in the crystal lattices of the grains are evicted by the light and return to their normal atomic sites. **b**, When grains are buried and hidden from sunlight, environmental radiation causes electrons to leave these sites and be captured by the traps. **c**, If the grains are collected (concealing them from daylight), prepared in the laboratory and illuminated by infrared or visible (green or blue) light, emptying of the traps gives rise to optically stimulated luminescence (OSL). This is amplified by a photomultiplier tube and measured using a photon-counting system. The past radiation dose — from which the burial time of the grains is determined — is estimated as the equivalent dose of laboratory radiation needed to produce an OSL signal of the same intensity[1]. The OSL signals are separated from unwanted emissions and light from the stimulation beam using filters.

questions on topics ranging from landscape dynamics, climate change and soil development to human evolution and dispersal over the past few hundred millennia, as well as more recent archaeological events[3–7]. For example, optical dating has revealed that symbolic markings, personal ornaments and innovative technology associated with early modern humans appeared more than 70,000 years ago in southern Africa, and were widespread across the region by 60,000 years ago[14–16] — some 15 millennia before modern humans entered Europe. The technique has also had a key role in establishing that humans had arrived in Australia by around 50,000 years ago[17] and that the last of the 'megafauna' (the giant marsupials, reptiles and flightless birds that once roamed the continent) perished soon after[18], during a period of increasing aridity but preceding a protracted phase of much drier climate[4,17].

Methodological and instrumental developments continue to drive advances in optical dating. Many quartz grains have physical properties that are ill-suited to SAR procedures, and two further possible complications are insufficient bleaching of grains before deposition and mixing of sediments after burial. Measurements of individual grains — the fundamental unit of analysis in optical dating — allow each of these factors to be investigated for sand-sized grains using SAR procedures[12]. This has helped to improve the accuracy of optical ages by reducing the uncertainties inherent in measuring composite OSL signals from multiple grains[7,15].

Nonetheless, other constraints on optical dating remain, and will keep researchers busy searching for solutions. A key limitation is the time range, which is governed by the maximum number of electrons that can be caught in light-sensitive traps and by the long-term stability of the traps at environmental temperatures. Optical-dating applications have largely been restricted to the past 200,000 years, and efforts to push the maximum limit beyond the 800,000-year timespan investigated by Huntley et al.[1] have mostly ended in disappointment. But new vistas are opening up, with the recent identification of longer-range optical-dating signals in quartz and feldspar[5–7]. If these are confirmed as reliable chronometers, then optical dating of major events in Earth and human history during the Early Pleistocene — the period from about 2.6 million to 0.8 million years ago — has a bright future.

Developments are also afoot to map the distribution of optical ages for individual grains on the cut surfaces of intact sediments and artefacts[7]. The ability to acquire such spatially resolved ages would be an advance over the current practice of disaggregating samples to extract grains for OSL measurements, which results in the loss of valuable contextual information. The fundamental insights obtained would be on a par with those gained from

single-crystal dating in other branches of geology and in single-cell analysis in biology.

New frontiers for optical dating also include the use of OSL signals to investigate the long-term exhumation of landscapes and evolution of mountain ranges[19], and the in situ dating of minerals on Mars using robotic devices, which would propel optical dating into space[20]. These applications are extremely challenging, but if the past 30 years of progress are any guide, we can expect optical dating to illuminate much more of the history of this planet — and perhaps that of others — before we celebrate its fiftieth anniversary. ∎

Richard G. Roberts is at the Centre for Archaeological Science, School of Earth and Environmental Sciences, University of Wollongong, Wollongong, New South Wales 2522, Australia. **Olav B. Lian** is in the Department of Geography and the Environment, University of the Fraser Valley, Abbotsford, British Columbia V2S 7M8, Canada.
e-mails: rgrob@uow.edu.au; olav.lian@ufv.ca

1. Huntley, D. J., Godfrey-Smith, D. I. & Thewalt, M. L. W. Nature **313,** 105–107 (1985).
2. Aitken, M. J. An Introduction to Optical Dating (Oxford Univ. Press, 1998).
3. Lian, O. B. & Roberts, R. G. Quat. Sci. Rev. **25,** 2449–2468 (2006).
4. Prescott, J. R. & Robertson, G. B. Aust. J. Earth Sci. **55,** 997–1007 (2008).
5. Rhodes, E. J. Annu. Rev. Earth Planet. Sci. **39,** 461–488 (2011).
6. Liritzis, I. et al. Luminescence Dating in Archaeology, Anthropology, and Geoarchaeology: An Overview (Springer, 2013).
7. Roberts, R. G. et al. J. Archaeol. Sci. **56,** 41–60 (2015).
8. Wintle, A. G. & Huntley, D. J. Nature **279,** 710–712 (1979).
9. Smith, B. W., Aitken, M. J., Rhodes, E. J., Robinson, P. D. & Geldard, D. M. Radiat. Protect. Dosim. **17,** 229–233 (1986).
10. Hütt, G., Jaek, I. & Tchonka, J. Quat. Sci. Rev. **7,** 381–385 (1988).
11. Murray, A. S. & Roberts, R. G. Radiat. Meas. **29,** 503–515 (1998).
12. Galbraith, R. F., Roberts, R. G., Laslett, G. M., Yoshida, H. & Olley, J. M. Archaeometry **41,** 339–364 (1999).
13. Bøtter-Jensen, L., McKeever, S. W. S. & Wintle, A. G. Optically Stimulated Luminescence Dosimetry (Elsevier, 2003).
14. Henshilwood, C. S. et al. Science **295,** 1278–1280 (2002).
15. Jacobs, Z. et al. Science **322,** 733–735 (2008).
16. Brown, K. S. et al. Nature **491,** 590–593 (2012).
17. Bowler, J. M. et al. Nature **421,** 837–840 (2003).
18. Roberts, R. G. et al. Science **292,** 1888–1892 (2001).
19. Herman, F., Rhodes, E. J., Braun, J. & Heiniger, L. Earth Planet. Sci. Lett. **297,** 183–189 (2010).
20. Kalchgruber, R., Blair, M. W., McKeever, S. W. S., Benton, E. R. & Reust, D. K. Planet. Space Sci. **55,** 2203–2217 (2007).

STRUCTURAL BIOLOGY

Pain–sensing TRPA1 channel resolved

The TRPA1 ion channel activates pain pathways in response to noxious compounds. The structure of TRPA1 has now been solved, providing insight into how it functions. SEE ARTICLE P.511

DAVID E. CLAPHAM

Rooted in place, many plants resort to chemical warfare to survive predation by bacteria, fungi, insects and mammalian herbivores. They produce pungent natural chemicals, such as capsaicin, which makes chilli peppers 'hot', and the thiosulfinates that make onion chopping a tear-jerker. Reactive chemicals in onions, wasabi and related spices activate the ion channel TRPA1, a relative of the capsaicin receptor TRPV1. On page 511 of this issue, Paulsen et al.[1] follow up their description of TRPV1 (refs 2, 3). Using electron cryo-microscopy techniques, they define the full-length, single-particle structure of TRPA1 to around 4 ångströms, a level of resolution that reveals its general features.

TRP channels are found in almost all cell types in eukaryotes (organisms that

include plants, animals and fungi). There are 27 human TRP-channel genes, which mostly encode weakly selective cation channels that enable ion flux across membranes in response to the binding of extracellular or intracellular ligands, or to changes in temperature or membrane voltage. On opening, TRP channels reduce the voltage across membranes and enable cations such as calcium ions (Ca^{2+}) to flow into the cytoplasm[4].

TRPA1 is found in the plasma membranes of pain-detecting sensory nerves[4]. It activates pain pathways that trigger avoidance behaviours and pathways that promote longer-lasting biological responses, such as inflammation[5]. Blocking TRPA1 function is therefore a promising strategy for treating pain. Pungent agents from wasabi, and other TRPA1 triggers, are known to be electrophiles, activating the channel by forming covalent bonds with specific cysteine or lysine
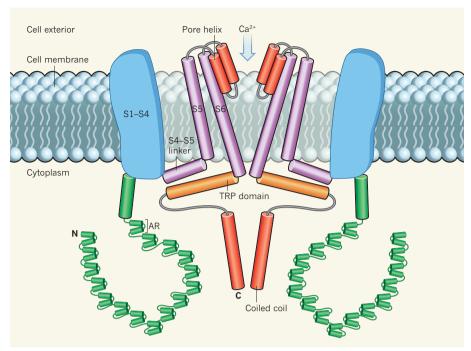
Figure 1 | The structure of TRPA1. A schematic of key features of the ion channel TRPA1, which mediates the cellular influx of cations such as calcium ions (Ca^{2+}). Two TRPA1 subunits are shown, although the channel is comprised of four. Paulsen et al.[1] report that each subunit contains six membrane-spanning α-helical domains, S1–S6. Two pore helices link S5 and S6 at the extracellular surface, where a constriction regulates influx of Ca^{2+}. The helical TRP domain is part of a second, lower constriction. Sixteen ankyrin repeats (ARs) at the amino-terminal end (N) of the subunit cover a carboxy-terminal (C) coiled-coil structure, providing a large cytoplasmic surface for interactions with noxious agents. It is probable that the molecular interactions of ligands with ARs lead to conformational changes, conveyed through the S4–S5 linker structure, that open the channel. The lower constriction is closed to ions in the authors' structure, and may be in either a closed or desensitized state.

amino-acid residues. However, the overall organization of the channel has been a mystery, frustrating efforts to understand exactly how ligand binding activates TRPA1.

Ion channels that open in response to ligand binding or voltage changes typically share the same primary architecture, in which each subunit of the channel contains six membrane-spanning α-helical domains, S1–S6 (ref. 6). Paulsen et al. report that, like archetypal voltage-gated potassium channels[6], four TRPA1 subunits come together to form a homotetrameric ion channel, with domains S5 and S6 of each subunit contributing to a shared ion-conducting pore. In highly voltage-sensitive channels, the S1–S4 domain is a charged adaptor, which enables voltage changes across the membrane to open and close (gate) the pore[6]. TRP channels are less voltage-sensitive than some other channels, and their S1–S4 domain probably translates the energy of ligand binding into this gating movement.

Paulsen and co-workers show that TRPA1 channels, like other ion channels, have two primary constrictions, one at either end of the channel's ion-conducting pathway (Fig. 1). Similar to bacterial voltage-gated sodium channels[7,8], each subunit of the channel's outer pore domain contains two short 'pore helices', which point into the ion-conduction pathway

— a difference from the typical solitary pore helix found in other channels. These helices slope steeply down to the first constriction and ion-selectivity filter, which is big enough to allow partially hydrated Ca^{2+} ions to pass through. The authors find that one type of TRPA1 antagonist binds to the channel close to S5, S6 and the first pore helix, which could explain its ability to inhibit gating.

The second constriction is formed by two hydrophobic amino-acid residues from each subunit. It is lower down in the membrane, close to the cytoplasmic face of the channel, and the authors seem to have captured it in a closed conformation.

For channel aficionados, this is a familiar story, so what is different about the TRPA1 structure? Paulsen and colleagues find that around 80% of TRPA1's mass is outside the channel's core, in the amino- and carboxy-terminal domains. A coiled-coil domain in the C terminus forms the cytoplasmic stalk of the channel, and is surrounded by 'ankyrin repeats', which it is speculated[9] contain the cysteine residues targeted by electrophilic TRPA1 activators. An ankyrin repeat is a motif of 33 amino acids that forms two α-helices separated by loops. They are one of the most common, evolutionarily conserved structural motifs, and generally form protein-interaction domains[10].

Paulsen and co-workers' three-dimensional reconstruction and modelling experiments indicate that each of the 4 subunits of TRPA1 has at least 16 ankyrin repeats, 5 of which are well resolved in the authors' structure and 11 of which their models suggest form a propeller-like structure, resembling the backs of four armadillos, suspended below the membrane (see Fig. 1c of the paper[1]). The researchers propose that this structure regulates gating of the pore.

All TRP channels contain a TRP domain — a helix at the base of S6. In Paulsen and co-workers' structure, this helix runs perpendicular to the pore helices, parallel to the membrane and suspended above the ankyrin repeats, where it might contribute to regulation of the lower pore. Perhaps most interesting is the coiled-coil stalk, which seems to mediate bundling of the four subunits through interactions between predicted α-helices at the base of the channel. This is quite different from TRPV1, which has a splayed base formed by six ankyrin repeats, one of which mediates subunit interactions by contacting a three-stranded β-sheet structure on an adjacent subunit[2,3].

Negatively charged phosphatidylinositol lipids such as $PtdIns(4,5)P_2$ are stationed at membrane surfaces, and are involved in signal-transduction cascades that regulate the activity of many ion channels and transporters[11]. Soluble polyphosphate molecules, such as inositol hexakisphosphate ($InsP_6$), are less well known as channel regulators. Paulsen et al. found that the presence of $InsP_6$ was necessary for channel formation, and provide evidence that it binds directly to the channel, close to the coiled coil. $InsP_6$ is negatively charged, and the authors report charge–charge interactions with four adjacent, positively charged amino acids in the coiled coil. Thus, $InsP_6$ seems to be a helper molecule, required for successful channel formation or stabilization.

Defining the structure of TRPA1, like the previous achievement with TRPV1, is a milestone in TRP-channel biology. These two structures suggest that the basic architecture of TRPs is broadly similar to that of voltage-gated potassium channels — they assemble as tetramers surrounding variably sized and charged gates. Although the latest study does not tell us how electrophiles and other noxious compounds open the TRPA1 channel, it suggests that they evoke interactions through the large surface area of the ankyrin repeats. In the future, more-detailed structures of TRPA1 in different conformations will reveal regulatory features, such as why the channel becomes sensitized and desensitized to calcium, and, perhaps more importantly, how channel function can be blocked to treat asthma, inflammation and pain. ■

David E. Clapham is at the Howard Hughes Medical Institute, Department of

Cardiology, Boston Children's Hospital, and
the Department of Neurobiology, Harvard
Medical School, Boston, Massachusetts
02115, USA.
e-mail: dclapham@enders.tch.harvard.edu

1. Paulsen, C. E., Armache, J.-P., Gao, Y., Cheng, Y. &
   Julius, D. Nature 520, 511–517 (2015).
2. Liao, M., Cao, E., Julius, D. & Cheng, Y. Nature 504,
   107–112 (2013).
3. Cao, E., Liao, M., Cheng, Y. & Julius, D. Nature 504,
   113–118 (2013).
4. Ramsey, I. S., Delling, M. & Clapham, D. E.
   Annu. Rev. Physiol. 68, 619–647 (2006).
5. Julius, D. Annu. Rev. Cell Dev. Biol. 29, 355–384
   (2013).
6. Jiang, Y. et al. Nature 423, 33–41 (2003).
7. Zhang, X. et al. Nature 486, 130–134 (2012).
8. Payandeh, J., Gamal El-Din, T. M., Scheuer, T., Zheng, N.
   & Catterall, W. A. Nature 486, 135–139 (2012).
9. Nilius, B., Appendino, G. & Owsianik, G. Pflügers
   Arch. 464, 425–458 (2012).
10. Herbert, M. H., Squire, C. J. & Mercer, A. A. Viruses 7,
   709–738 (2015).
11. Suh, B.-C. & Hille, B. Annu. Rev. Biophys. 37,
   175–195 (2008).

OPTOMECHANICS

Listening to quantum grains of sound

An optomechanical device has allowed quanta, or 'grains', of mechanical vibration to be counted by optical means. The system may open up new possibilities in acoustics and thermal engineering. SEE LETTER P.522

IVAN FAVERO

According to quantum theory, all forms of energy come in tiny 'grains' called quanta. This quantum granularity may or may not be discernible, depending on the instruments available and the nature of the energy. The quanta are photons in the case of electromagnetic waves such as light, and phonons for mechanical vibrations in solids. Phonons are ubiquitous in condensed-matter systems, in which they underlie the transport of sound and heat. They also govern the performance of electronic and optical devices, and play a central part in conventional superconductivity. However, whereas photons can be detected at the single-quantum level using today's optical receivers, there is a lack of versatile instruments and techniques to measure phonons at the same level of precision. The scientific and technological stakes are high, given the many potential applications of such systems. On page 522 of this issue, Cohen et al.[1] describe a promising step towards building an optical single-phonon detector.

The researchers used an optomechanical device in which photons from an external light source are scattered by phonons associated with the system's mechanical vibration. The phonons are detected by collecting the scattered photons in single-photon detectors. The experiment builds on developments in the field of optomechanics, in which the force of light, circulating in a type of light trap called an optical cavity, is used to control and monitor the mechanical motion of tiny objects such as mirrors, membranes or flexible wires[2,3]. Although light generally has a feeble mechanical effect on large bodies, it dictates the dynamics of such minute optomechanical systems. Progress in the field over the past five years has involved confining both photons and phonons in a submicrometre-sized volume to achieve strong coupling between light and mechanical motion[4,5].
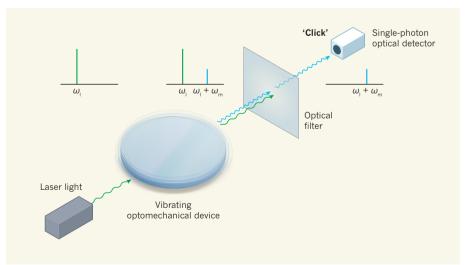
Cohen and colleagues' optomechanical device is a nanostructured semiconductor crystal that yields such strong coupling. As a result, laser light that is incident on, and scattered from, the optomechanical crystal is efficiently modulated by the crystal's mechanical vibration. The scattered light acquires blue and red spectral sidebands, respectively above and below the laser's frequency, that are analogous to the 'satellite lines' observed in the Raman scattering of photons from a material. If the material absorbs energy during the scattering process, the scattered photon has a lower energy than the incident photon (red Stokes line); if the material loses energy, the scattered photon has a higher energy (blue anti-Stokes line). The shift in energy provides information about the vibrational or rotational modes of motion of the material's constituents. In a sense, optomechanical devices are simply highly engineered Raman systems.

By appropriately tuning the laser's frequency to the optical resonance of the optomechanical crystal, the system's natural optical vibration frequency, Cohen et al. could finely adjust the amplitude of the sidebands. For example, they suppressed the amplitude of the red sideband and selectively enhanced the generation of scattered blue photons (Fig. 1). By using optical filtering, the researchers then suppressed scattered photons that had the same energy as the incident laser photons. In this way, only the blue photons produced 'clicks' on the detector, revealing the presence of phonons in the device. Owing to the strong optomechanical interaction, the optical measurement attained sufficient sensitivity to resolve individual phonons, and enabled their precise counting in the system's vibrational state.

Although the present experiments still fall short of actually generating and measuring vibrational states containing a single phonon, the authors' phonon-counting technique already

Figure 1 | Optical phonon counting. Cohen et al.[1] shone laser light of frequency ω_l on an optomechanical device. The light's photons interacted strongly with phonons (quanta of vibration associated with the device's mechanical motion) of frequency ω_m, such that the output light acquired a blue spectral sideband at $\omega_l + \omega_m$ or a red spectral sideband at $\omega_l - \omega_m$. (For simplicity, only the former case is shown.) The output light was sent through an optical filter that suppressed photons of frequency ω_l but let through blue photons, which were directed to a single-photon optical detector. Photon 'clicks' at the detector indicated the presence of phonons in the system.

allows measurement of the statistical properties of the device's vibrational motion. The authors applied the concept of intensity correlations to their experimental set-up. Such correlations were used by Robert Hanbury Brown and Richard Twiss nearly 60 years ago to detect correlations between photons emitted by distant stars[6]. Cohen *et al.* measured these correlations in the arrival of photons at the detectors, thereby probing phonon correlations in the optomechanical device. Using this technique, they could directly observe a transition in the statistical behaviour of phonons as the system underwent a change from a purely thermal, random state of vibrational motion to a coherent, more-ordered one, which was reached above a certain threshold of power of the incident laser light[7,8]. In other words, Cohen and colleagues observed the phononic analogue of the 'lasing' transition that

enables lasers to emit coherent light — light that is made up of waves that have the same wavelength and are in step with each other.

Finally, and looking ahead, it should be noted that, in the realm of quantum physics, measuring is also acting. The very act of measuring a system may alter its state. Therefore, future optomechanical experiments operating in the single-phonon counting regime could be used to generate complex quantum states of phonons. Given the rapid pace at which the field of optomechanics is advancing, this point might be reached in the not-too-distant future. Such quantum optomechanical control of matter could allow researchers both to test the fundamental principles of quantum mechanics and to venture into new applications in acoustics, thermal management and electrical-conductivity engineering. ■

Ivan Favero *is at the Laboratoire Matériaux et Phénomènes Quantiques, Université Paris Diderot, CNRS, UMR 7162, F-75205 Paris Cedex 13, France.*
e-mail: ivan.favero@univ-paris-diderot.fr

1. Cohen, J. D. *et al. Nature* **520,** 522–525 (2015).
2. Favero, I. & Karrai, K. *Nature Photon.* **3,** 201–205 (2009).
3. Aspelmeyer, M., Kippenberg, T. J. & Marquardt, F. *Rev. Mod. Phys.* **86,** 1391–1452 (2014).
4. Chan, J., Safavi-Naeini, A. H., Hill, J. T., Meenehan, S. & Painter, O. *Appl. Phys. Lett.* **101,** 081115 (2012).
5. Baker, C. *et al. Opt. Express* **22,** 14072–14086 (2014).
6. Hanbury Brown, R. & Twiss, R. Q. *Nature* **178,** 1046–1048 (1956).
7. Hohberger, C. & Karrai, K. *Proc. 4th IEEE Conf. Nanotechnol.* 419–421 (IEEE, 2004).
8. Carmon, T., Rokhsari, H., Yang, L., Kippenberg, T. J. & Vahala, K. J. *Phys. Rev. Lett.* **94,** 223902 (2005).

EARTH SCIENCE

Landscape inversion by stream piracy

A model suggests that active deformation in mountains causes river networks to constantly reorganize, providing an explanation for the paradoxical formation of almost flat surfaces high in craggy mountain ranges. SEE LETTER P.526

JÉRÔME LAVÉ
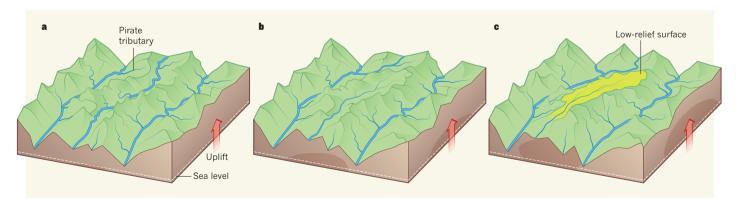
For more than a century, Earth scientists' curiosity has been piqued by the existence of areas with low topographic relief, some nearly flat, perched high in rugged mountain ranges. One common explanation[1] posits that these surfaces are relics of large peneplains — low-relief features formed as the ultimate result

of fluvial erosion — that were once close to sea level before being uplifted by mantle convection or plate tectonics and then dissected by rivers or streams. On page 526 of this issue, Yang *et al.*[2] propose a very different mechanism, whereby low-relief surfaces in mountain landscapes form transiently as a result of the dynamic reorganization of river networks.

One of Earth science's greatest challenges is

to document vertical movements of the crust at geological timescales. Step-like landforms called terraces are commonly used as passive markers of small-scale deformations that occurred during the past million years or less, but it is usually impossible to use such markers for large uplifted regions and events that occurred over longer time periods. Geologists are therefore forced to use other palaeoaltimetric methods[3] that generally have large uncertainties and are difficult to implement over large regions.

The lack of reliable tools is particularly problematic for studies of regional uplift caused by deep-seated geodynamic processes, or when investigating the upheaval of wide orogenic plateaus (which form as a result of colliding tectonic plates). In these cases, it is tempting to find other passive markers, such as the dissected and uplifted remnants of peneplains that are assumed to have once been nearly horizontal and close to sea level. By interpreting low-relief surfaces perched across

Figure 1 | Proposed origin of low-relief surfaces at high elevation. Yang *et al.*[2] suggest that when tectonic plates collide, the resulting large-scale deformation of the crust and upper mantle triggers permanent reorganization of river networks. **a,** In this illustration, a mountainous region is subjected to uplift, and a 'pirate' tributary of the left-hand river is indicated. **b,** Over time, the upstream part of the central river is captured by the pirate tributary,

causing a sudden decrease in the central river's stream power and its ability to incise through bedrock. Sustained tectonic uplift is no longer equilibrated by fluvial erosion, leading to uplift of the disconnected valley, with continuous erosion of the hillslope around the valley lowering the relief. **c,** A low-relief surface at high elevation emerges and may survive for some time before being degraded or captured by streams eroding inwards from its outer perimeter.

the southeast Tibetan plateau in this way, a broad south-dipping slope of the plateau surface has been inferred[4], which was proposed as evidence of differential thickening of the underlying crust. Aspects of this interpretation have been questioned[5,6], but a two-stage mechanism has always been assumed, in which the formation of low-relief surfaces precedes river dissection.

Yang et al. challenge this idea. They first propose a new description of the fluvial network that drains southeastern Tibet, by using a modified metric[7] of the stream power equation, a widely used model in which fluvial incision into bedrock primarily depends on river slope and discharge. This model predicts that when regional uplift increases (as expected in the dissected-peneplain scenario), the profiles of all the major rivers and tributaries will display a change of slope (a 'knickpoint') at similar elevations[8] if the geometry of the fluvial network is fixed. By examining several regions in southeastern Tibet, the researchers show that the main tributaries do not display such a regular pattern. Instead, knickpoint elevations are widely scattered.

The authors then show that the long profiles of tributaries that drain low-relief surfaces are systematically shallower than expected, whereas those of tributaries that drain the slopes surrounding the low-relief surfaces are steeper. The authors interpret these peculiar features[9] as evidence of recent drainage captures — the diversion of headwater regions of main rivers to nearby tributaries — and dynamic reorganization of the river network (Fig. 1). The low-relief surfaces therefore cannot be relict landscapes, and must instead correspond to areas that formerly had normal relief, but where the main stream has lost its power to incise bedrock because its headwater has been captured. If subjected to sustained uplift, such areas would gain elevation, and their local relief would be smoothed down because of erosion of the surrounding hill slope. The process is a kind of topographic inversion, because formerly incised valley bottoms and rugged topographies end up as flatter surfaces at high elevations.

To simulate this mechanism, the authors built a computational model in which a section of crust is squeezed between two rigid plates while undergoing constant thickening — a situation that may have occurred in southeastern Tibet[10]. This model and the associated video (see the paper's Supplementary Information[2]) wonderfully illustrate how crustal shrinkage reduces the overall drainage area that feeds the rivers, causing continuous reorganization in which 'victim' rivers lose their upstream area to 'pirate' river networks. Crucially, they demonstrate how stream piracy in this deforming setting is a self-sustained or cascading mechanism. Once a network has lost part of its drainage area, its ability to incise the crust decreases, its elevation above surrounding

major rivers increases, and it becomes easier for pirate networks to capture even more of its drainage area.

Yang and colleagues' decision to reject the classical explanation of the low-relief surfaces in southeastern Tibet merits some discussion. First, part of the observed scattering of knickpoint elevations might result from river-incision behaviour that is not encapsulated in the simplified stream power model used by the authors, from local variations of tectonics, or from the initial topography of the raised low-relief surface. Second, the authors' model does not easily apply to the northern part of the studied region (north of 30° N), where there have been low rates of erosion during the past 50 million years[11] and where reduced river incision has probably limited the reorganization of fluvial networks; low-relief surfaces in this region represent more than half of the landscape, whereas the model predicts a much smaller fraction.

Future modelling should investigate the roles of horizontal strain and vertical uplift in the dynamics of river capture. Would the model have led to such a dynamic stream reorganization and production of low-relief areas if the authors had considered much lower finite erosion[11] and the present north–south flow deformation[12] in southeastern Tibet, which is in sharp contrast to the east–west shortening simulated in the model?

In any case, the process proposed by Yang et al. changes our thinking about the genesis of low-relief surfaces in mountainous areas, and will reignite debate about their origin and use as uplift markers in other orogenic settings, such as the Pyrenees[13] or the Eastern Andes cordillera[14]. The multicapture scenario will also alter the way we look at the evolution of river networks in many landscapes, and, as a corollary, affect interpretations of sedimentary archives that have recorded past erosion of regions undergoing deformation. Finally, if stream reorganization is highly sensitive to horizontal and vertical tectonics, it highlights both the richness and the complexities provided by river-network geometry for unravelling the tectonic history of orogenic features[10]. ∎

Jérôme Lavé is at the Centre de Recherches Pétrographiques et Géochimiques, CNRS, Vandoeuvre les Nancy 54501, France.
e-mail: jlave@crpg.cnrs-nancy.fr

1. Davis, W. M. Ann. Assoc. Am. Geogr. 1, 21–83 (1911).
2. Yang, R., Willett, S. D. & Goren, L. Nature 520, 526–529 (2015).
3. Kohn, M. J. (ed.) Paleoaltimetry: Geochemical and Thermodynamic Approaches (Mineral. Soc. Am., 2007).
4. Clark, M. K. et al. J. Geophys. Res. 111, F03002 (2006).
5. Hoke, G. D., Liu-Zeng, J., Hren, M. T., Wissink, G. K. & Garzione, C. N. Earth Planet. Sci. Lett. 394, 270–278 (2014).
6. Liu-Zeng, J., Tapponnier, P., Gaudemer, Y. & Ding, L. J. Geophys. Res. 113, F04018 (2008).
7. Perron, J. T. & Royden, L. Earth Surf. Process.

Landforms **38**, 570–576 (2013).
8. Niemann, J. D., Gasparini, N. M., Tucker, G. E. & Bras, R. L. *Earth Surf. Process. Landforms* **26**, 1317–1332 (2001).
9. Willett, S. D., McCoy, S. W., Perron, J. T., Goren, L. &

Chen, C.-Y. *Science* **343**, 1248765 (2014).
10. Hallet, B. & Molnar, P. *J. Geophys. Res.* **106**, 13697–13709 (2001).
11. Clark, M. K. *et al. Geology* **33**, 525–528 (2005).
12. Shen, Z.-K., Lü, J., Wang, M. & Bürgmann, R.

J. Geophys. Res. **110**, B11409 (2005).
13. Gunnell, Y. *et al. Earth Planet. Sci. Lett.* **278**, 208–218 (2009).
14. Gubbels, T. L., Isacks, B. L. & Farrar, E. *Geology* **21**, 695–698 (1993).

SENSORY SYSTEMS

The yin and yang of cortical oxytocin

Female mice can learn to respond to distress calls from young mice — an ability that has now been found to be improved through signalling by the hormone oxytocin in the left auditory cortex of the brain. SEE ARTICLE P.499

ROBERT C. LIU

When newborns cry in the night, they capture the attention of sleepy parents, who become suddenly alert, ready to comfort their babies. This response might seem like instinct, but our ability to recognize social cues from infants is heavily shaped by experience[1,2]. Mothers and fathers of many species, including humans, learn over time to recognize the cries of their own babies[3]. In this issue, Marlin *et al.*[4] (page 499) investigate how the brain learns this information. They report that oxytocin, a well-studied hormone released in the brain in social situations, acts in an unexpected manner to help to create memories of infant cues.

Oxytocin is a neuropeptide molecule produced in the brain's hypothalamus. It acts in both the peripheral and central nervous systems, exerting prosocial effects on behaviour by promoting pair bonding, parental care, social reward, and attention to and memory of social cues[5–7]. But how oxytocin actually acts on neurons to affect social behaviours is only just beginning to be explored[8,9], with much still unknown about the mechanisms by which it influences sensory-information processing and memory in social contexts.
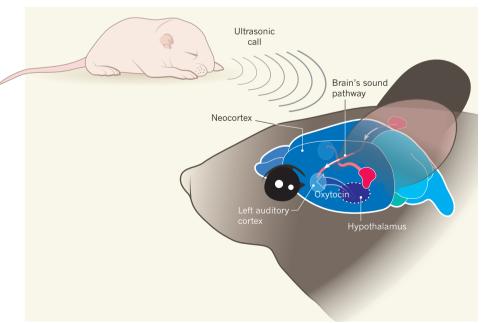
To address this gap in knowledge, Marlin and colleagues studied how female mice react to the ultrasonic cries made by pups. Mothers and virgin 'nanny' mice (which have experience of caring for pups) recognize the cries

as distress calls and respond by approaching calling pups, picking them up and carrying them to the nest. By contrast, inexperienced (naive) mice fail to recognize the cue and do not retrieve pups. There is evidence[10] that memories of the pups' cries in maternal mice are linked to the activity of neurons in the auditory cortex of the brain — part of the neocortex, which controls higher brain functions, including sensory perception (Fig. 1). Marlin *et al.* used pharmacological inactivation to demonstrate that, without left-auditory-cortical activity, pup retrieval is severely impaired in most experienced female mice. Inactivating the right auditory cortex had little effect, adding neuronal evidence to other work[11] suggesting that control of communication processing in mice is dominated by one side of the brain (lateralized), much as it is in other species.

Systemically delivered oxytocin facilitates maternal behaviours such as pup retrieval[2,7], but it has been presumed to act on evolutionarily conserved circuits for maternal responsiveness that are located in subcortical brain regions beneath the neocortex. Oxytocin-receptor proteins have previously been discovered in the prefrontal region of the neocortex[9], but Marlin *et al.* found that both oxytocin receptors and projections from hypothalamic oxytocin-producing neurons are present in the auditory cortex of mice, with the former (but not the latter) being more numerous on the left side than on the right.

Amazingly, naive virgins that had been injected with oxytocin in the left auditory cortex began retrieving pups earlier than counterparts that received saline. It remains to be seen whether oxytocin's cortical effect on pup retrieval is lateralized, because the authors did not repeat the experiment in the right auditory cortex. Blocking oxytocin-receptor activation in the left auditory cortex of experienced retrievers did not impair performance, which perhaps indicates that oxytocin in the auditory cortex facilitates learning about calls rather than maintaining memories of them. If so, blocking auditory-cortical oxytocin receptors during endogenous systemic oxytocin release in naive mice should prevent them from learning pup-retrieval behaviours.

What is the role of oxytocin in the auditory cortex? In technically challenging experiments, Marlin and colleagues analysed the neuronal inputs to auditory-cortical neurons, and showed that oxytocin can alter the balance between inhibitory and excitatory inputs, which are both activated by pup calls — they are the yin and yang of neuronal signalling. The team found that inhibitory and

Figure 1 | From cry to cortex. Ultrasonic distress calls made by young mice act as social cues that elicit a maternal response in females that have experience of caring for pups. Memories associated with this cue seem to be linked to neurons in the auditory cortex, part of the neocortex region of the brain. The hormone oxytocin, which is produced in the hypothalamus, can be released in the neocortex. Marlin *et al.*[4] report that oxytocin can act directly in the left auditory cortex to facilitate the behavioural response to the cries of pups, which are relayed to this brain region through a sound pathway.

Landforms **38**, 570–576 (2013).
8. Niemann, J. D., Gasparini, N. M., Tucker, G. E. & Bras, R. L. *Earth Surf. Process. Landforms* **26**, 1317–1332 (2001).
9. Willett, S. D., McCoy, S. W., Perron, J. T., Goren, L. &

Chen, C.-Y. *Science* **343**, 1248765 (2014).
10.Hallet, B. & Molnar, P. *J. Geophys. Res.* **106**, 13697–13709 (2001).
11.Clark, M. K. *et al. Geology* **33**, 525–528 (2005).
12.Shen, Z.-K., Lü, J., Wang, M. & Bürgmann, R.

J. Geophys. Res. **110**, B11409 (2005).
13.Gunnell, Y. *et al. Earth Planet. Sci. Lett.* **278**, 208–218 (2009).
14.Gubbels, T. L., Isacks, B. L. & Farrar, E. *Geology* **21**, 695–698 (1993).

SENSORY SYSTEMS

The yin and yang of cortical oxytocin

Female mice can learn to respond to distress calls from young mice — an ability that has now been found to be improved through signalling by the hormone oxytocin in the left auditory cortex of the brain. SEE ARTICLE P.499

ROBERT C. LIU

When newborns cry in the night, they capture the attention of sleepy parents, who become suddenly alert, ready to comfort their babies. This response might seem like instinct, but our ability to recognize social cues from infants is heavily shaped by experience[1,2]. Mothers and fathers of many species, including humans, learn over time to recognize the cries of their own babies[3]. In this issue, Marlin *et al.*[4] (page 499) investigate how the brain learns this information. They report that oxytocin, a well-studied hormone released in the brain in social situations, acts in an unexpected manner to help to create memories of infant cues.

Oxytocin is a neuropeptide molecule produced in the brain's hypothalamus. It acts in both the peripheral and central nervous systems, exerting prosocial effects on behaviour by promoting pair bonding, parental care, social reward, and attention to and memory of social cues[5–7]. But how oxytocin actually acts on neurons to affect social behaviours is only just beginning to be explored[8,9], with much still unknown about the mechanisms by which it influences sensory-information processing and memory in social contexts.

To address this gap in knowledge, Marlin and colleagues studied how female mice react to the ultrasonic cries made by pups. Mothers and virgin 'nanny' mice (which have experience of caring for pups) recognize the cries

as distress calls and respond by approaching calling pups, picking them up and carrying them to the nest. By contrast, inexperienced (naive) mice fail to recognize the cue and do not retrieve pups. There is evidence[10] that memories of the pups' cries in maternal mice are linked to the activity of neurons in the auditory cortex of the brain — part of the neocortex, which controls higher brain functions, including sensory perception (Fig. 1). Marlin *et al.* used pharmacological inactivation to demonstrate that, without left-auditory-cortical activity, pup retrieval is severely impaired in most experienced female mice. Inactivating the right auditory cortex had little effect, adding neuronal evidence to other work[11] suggesting that control of communication processing in mice is dominated by one side of the brain (lateralized), much as it is in other species.

Systemically delivered oxytocin facilitates maternal behaviours such as pup retrieval[2,7], but it has been presumed to act on evolutionarily conserved circuits for maternal responsiveness that are located in subcortical brain regions beneath the neocortex. Oxytocin-receptor proteins have previously been discovered in the prefrontal region of the neocortex[9], but Marlin *et al.* found that both oxytocin receptors and projections from hypothalamic oxytocin-producing neurons are present in the auditory cortex of mice, with the former (but not the latter) being more numerous on the left side than on the right.

Amazingly, naive virgins that had been injected with oxytocin in the left auditory cortex began retrieving pups earlier than counterparts that received saline. It remains to be seen whether oxytocin's cortical effect on pup retrieval is lateralized, because the authors did not repeat the experiment in the right auditory cortex. Blocking oxytocin-receptor activation in the left auditory cortex of experienced retrievers did not impair performance, which perhaps indicates that oxytocin in the auditory cortex facilitates learning about calls rather than maintaining memories of them. If so, blocking auditory-cortical oxytocin receptors during endogenous systemic oxytocin release in naive mice should prevent them from learning pup-retrieval behaviours.

What is the role of oxytocin in the auditory cortex? In technically challenging experiments, Marlin and colleagues analysed the neuronal inputs to auditory-cortical neurons, and showed that oxytocin can alter the balance between inhibitory and excitatory inputs, which are both activated by pup calls — they are the yin and yang of neuronal signalling. The team found that inhibitory and

Figure 1 | From cry to cortex. Ultrasonic distress calls made by young mice act as social cues that elicit a maternal response in females that have experience of caring for pups. Memories associated with this cue seem to be linked to neurons in the auditory cortex, part of the neocortex region of the brain. The hormone oxytocin, which is produced in the hypothalamus, can be released in the neocortex. Marlin *et al.*[4] report that oxytocin can act directly in the left auditory cortex to facilitate the behavioural response to the cries of pups, which are relayed to this brain region through a sound pathway.

excitatory inputs were not well balanced in naive females, whose neurons fired less consistently in response to calls than did those of their experienced counterparts. By applying oxytocin to the left auditory cortex, the authors could transiently weaken inhibitory currents in naive females. This disinhibited the excitatory response, resulting in a gradual balancing of yin and yang through strengthening of both types of input. The change led to more-robust signalling in response to calls, similar to that observed in the left auditory cortex of mothers and nannies. This neuronal effect was long-lasting, suggesting that it might provide a key mechanism for establishing memories of socially relevant sounds in the auditory cortex.

Marlin and colleagues' data show that the effect of oxytocin on behaviour varies substantially between individuals. Could this be related to natural variation in the oxytocin-receptor gene, affecting either the protein's function or how much of it is present in each individual[5,6]? Sound-evoked cortical responses and their susceptibility to oxytocin manipulations also vary from neuron to neuron. It will be important to understand whether this variation also arises from differences in oxytocin-receptor expression, or whether it can be explained by other differences between cortical neurons that have yet to be considered in depth. For example, could it be due to the types of cortical neuron that express oxytocin receptors, the sound features that cortical neurons respond to, or where in the auditory cortex the cry-responsive neurons are found?

This study supports recent hypotheses about how hormones and sensory experience interact to shape the function of the maternal cortex[12]. Such 'oxytocin–experience' interactions have been implicated in the recognition of social odours through subcortical neurons[6,13], but the current work breaks ground by showing that such interactions also act at the level of the sensory neocortex. The authors' findings suggest that oxytocin can behave in a similar manner to other neuromodulator systems that are involved in attention and learning, by instigating an enduring neuronal plasticity that increases responsiveness to social stimuli.

Human imaging studies have enabled us to expand our models of the neuronal circuits in the maternal brain that respond to infant cues, from primarily subcortical circuits to those that include neocortical regions[14]. Using mice alongside these studies opens the door to dissecting the mechanisms that underpin hormone–experience interactions in the neocortical areas that process infant cues. The potential relevance of these combined experimental strategies to investigating disorders such as postpartum depression and autism, in which the salience of social cues is diminished[6], adds a translational element to the promise of Marlin and colleagues' work. ∎

Robert C. Liu *is in the Department of Biology, Silvio O. Conte Center for Oxytocin and Social Cognition, Emory University, Atlanta, Georgia 30322, USA.*
e-mail: robert.liu@emory.edu

1. Fleming, A. S., O'Day, D. H. & Kraemer, G. W. *Neurosci. Biobehav. Rev.* **23**, 673–685 (1999).
2. Numan, M. & Insel, T. R. *The Neurobiology of Parental Behavior* (Springer, 2003).
3. Gustafsson, E., Levréro, F., Reby, D. & Mathevon, N. *Nature Commun.* **4**, 1698 (2013).
4. Marlin, B. J., Mitre, M., D'amour, J. A., Chao, M. V. & Froemke, R. C. *Nature* **520**, 499–504 (2015).
5. Donaldson, Z. R. & Young, L. J. *Science* **322**, 900–904 (2008).
6. Bartz, J. A., Zaki, J., Bolger, N. & Ochsner, K. N. *Trends Cogn. Sci.* **15**, 301–309 (2011).
7. Rilling, J. K. & Young, L. J. *Science* **345**, 771–776 (2014).
8. Dölen, G., Darvishzadeh, A., Huang, K. W. & Malenka, R. C. *Nature* **501**, 179–184 (2013).
9. Nakajima, M., Görlich, A. & Heintz, N. *Cell* **159**, 295–305 (2014).
10. Galindo-Leon, E. E., Lin, F. G. & Liu, R. C. *Neuron* **62**, 705–716 (2009).
11. Ehret, G. *Nature* **325**, 249–251 (1987).
12. Banerjee, S. B. & Liu, R. C. *Front. Neuroendocrinol.* **34**, 300–314 (2013).
13. Ferguson, J. N., Aldag, J. M., Insel, T. R. & Young, L. J. *J. Neurosci.* **21**, 8278–8285 (2001).
14. Lorberbaum, J. P. *et al. Biol. Psychiat.* **51**, 431–445 (2002).

REGENERATIVE BIOLOGY

Neuregulin 1 makes heart muscle

Three studies reveal that augmentation of a signalling pathway involving the growth factor neuregulin 1 and its receptor protein ERBB2 can promote the generation of muscle cells in zebrafish, mice and infant heart tissue.

KATHERINE E. YUTZEY

Whether mature heart-muscle cells called cardiomyocytes can proliferate to make new muscle has been an area of contention for many years. The debate has implications for the treatment of cardiovascular disease, because the loss of cardiomyocytes can lead to heart failure and death. Mammalian cardiomyocytes were thought to stop proliferating in the first few days after birth[1]. However, evidence has emerged to suggest that adult mouse and human cardiomyocytes do proliferate, albeit in very small numbers[2,3]. Now, three studies (published in *eLIFE*[4], *Nature Cell Biology*[5] and *Science Translational Medicine*[6]) report that the growth factor neuregulin 1 (Nrg1) promotes heart regeneration in zebrafish, and cardiomyocyte proliferation in mammalian hearts.

Cardiomyocytes in adult mice and humans are thought to renew at a rate of less than 1% per year[2,3]. By contrast, adult zebrafish can regenerate up to one-third of their heart muscle from existing cardiomyocytes in a few months after injury[7]. What factors mediate this regeneration? Gemberling *et al.*[4] found that Nrg1 induces cardiomyocyte proliferation and therefore heart regeneration in zebrafish by signalling through its co-receptor protein, Erbb2.

Turning to mammalian hearts, D'Uva *et al.*[5] examined whether Erbb2 is required for heart-muscle growth in mice just after birth. Pups older than one week had lower expression of Erbb2 than younger mice, indicating that expression of the receptor is downregulated
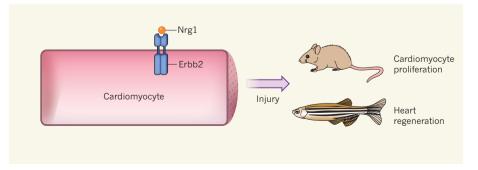

Figure 1 | Mending a broken heart. Three studies[4–6] of mice and zebrafish report that the growth factor neuregulin 1 (Nrg1) promotes the proliferation of heart-muscle cells called cardiomyocytes after injury. Nrg1 signals through the co-receptor protein Erbb2 on the membrane of cardiomyocytes to promote proliferation in mammals, and regeneration of heart muscle in fish. Note that the receptor is not shown to scale.

within a week of birth, at the same time as cardiomyocytes stop proliferating. The authors found that transient induction of an activated form of Erbb2 for 10–21 days in juvenile or adult cardiomyocytes prolonged the regenerative capacity of the hearts into adulthood or restored regenerative capacity, respectively. Expression of activated Erbb2 in mature cardiomyocytes enabled the cells to partially dedifferentiate to a less-specialized cell type after injury, resulting in disassembly of the muscle contractile apparatus and leading to proliferation in the hearts of juvenile and adult mice (Fig. 1).

In the third study, Polizzotti et al.[6] provide evidence to support D'Uva and colleagues' data. The authors injured the hearts of newborn mice through localized freezing. Control hearts showed scarring and had reduced function following injury, but heart function was preserved and cardiomyocyte proliferation was maintained in newborn mice treated with Nrg1. However, treatment started later than four days after birth did not improve heart function, possibly owing to the postnatal loss of Nrg1-receptor expression reported by D'Uva and colleagues. Together, then, these three studies support activation of the Nrg1–Erbb2 pathway as a way to prolong the ability of the newborn mouse heart to regenerate in the first few days after birth.

Whether there is a regenerative window in infants comparable to that in mice is unclear. Polizzotti and colleagues addressed this issue by adding NRG1 to diseased human heart biopsies that were cultured in vitro. The treatment promoted cardiomyocyte proliferation in biopsies taken from newborns, but the proliferative effect was significantly reduced by six months of age. Thus, there may be a narrow therapeutic window in which NRG1 treatment could improve the success of reconstructive surgery for infants born with severe heart defects.

Could NRG1 treatment also be a good way to treat heart disease in adults? D'Uva et al. demonstrated that transient induction of activated Erbb2 following a heart attack in adult mice led to improved heart function, reduced scarring and increased cardiomyocyte proliferation compared with mice that did not receive activated Erbb2. But the current studies[4,5] also demonstrate that, in both fish and mice, the unrestrained cardiomyocyte proliferation induced by Nrg1–Erbb2 signalling leads to enlargement and eventual failure of the heart. The correct level of signalling is therefore imperative — it may be beneficial to induce cardiomyocytes to proliferate with Nrg1, but it is equally important to shut off the pathway to maintain the appropriate number of cells for a healthy heart.

Activated ERBB2 was first identified in tumour cells, and expression of this mutated form of the receptor is an indicator of poor prognosis in breast cancer[8]. Therapies targeting activated ERBB2 are used to treat breast cancer, but heart-muscle damage is a known side effect and must be monitored in these patients[8]. Thus, another concern with using NRG1 therapy for heart disease is the possibility that this treatment might increase the risk of cancer. To minimize this risk, if activation of the NRG1–ERBB2 pathway is used to treat heart disease in the future, therapies should ideally specifically target cardiomyocytes, and the duration of treatments should be limited.

Other studies have tested the efficacy of Nrg1 treatment in adult mice following a heart attack, and have produced conflicting results. Although one study[9] reported that Nrg1 signalling stimulates proliferation of differentiated cardiomyocytes following injury, another[10] found no evidence of DNA synthesis, an indicator of proliferation, in Nrg1-treated hearts. The authors of the latter study speculated that Nrg1 might also protect the mammalian heart from injury, possibly by stimulating the growth of new blood vessels, which could explain the positive results of some studies in adult animals.

NRG1 is currently being tested in humans as a treatment for heart failure (https://clinical-trials.gov), but these trials are ongoing and it is too soon to tell whether the treatment is effective. The current studies suggest that NRG1 treatment might help the injured heart not only by supporting existing muscle, but also by promoting the production of new muscle. Both facets of NRG1 function should be considered in the development of treatments for cardiovascular disease based on regenerative pathways. ∎

Katherine E. Yutzey is in the Division of Molecular Cardiovascular Biology, Cincinnati Children's Medical Center, Cincinnati, Ohio 45229, USA.
e-mail: katherine.yutzey@cchmc.org

1. Soonpaa, M. H. & Field, L. J. Circ. Res. 83, 15–26 (1998).
2. Bergmann, O. et al. Science 324, 98–102 (2009).
3. Senyo, S. E. et al. Nature 493, 433–436 (2013).
4. Gemberling, M., Karra, R., Dickson, A. L. & Poss, K. D. eLIFE 4, e05871 (2015).
5. D'Uva, G. et al. Nature Cell Biol. http://dx.doi.org/10.1038/ncb3149 (2015).
6. Polizzotti, B. D. et al. Sci. Transl. Med. http://dx.doi.org/10.1126/scitranslmed.aaa5171 (2015).
7. Poss, K. D., Wilson, L. G. & Keating, M. T. Science 298, 2188–2190 (2002).
8. Cote, G. M., Sawyer, D. B. & Chabner, B. A. N. Engl. J. Med. 367, 2150–2153 (2012).
9. Bersell, K., Arab, S., Haring, B. & Kühn, B. Cell 138, 257–270 (2009).
10. Reuter, S., Soonpaa, M. H., Firulli, A. B., Chang, A. N. & Field, L. J. PLoS ONE 9, e115871 (2014).

ECOLOGY

Shared ancestry predicts disease levels

Ecological factors such as host density are important predictors of disease incidence. But another key determinant may be the evolutionary history and relatedness of the host community. SEE LETTER p.542

HELEN M. ALEXANDER

Understanding the dynamics of disease-causing microorganisms is central to human medicine and to improving the health of domesticated animals and crops. Disease is also an increasingly recognized part of ecological research, because pathogens can influence the coexistence of host species, the potential for non-native species to invade and ecosystem productivity[1,2]. However, predicting disease levels in an ecosystem is challenging. On page 542 of this issue, Parker et al.[3] argue that such predictions can be improved by understanding the evolutionary relationships of the host community — in ecological parlance, the assemblage of host species present at a location.

Communities of hosts and pathogens are complex. Starting simply, consider a single pathogenic species infecting a single host species. In this case, high host density is typically associated with increased disease — a relationship evident to any parent whose child brings a cold virus home from school. But what happens when there are multiple hosts? In some situations, increased host diversity might reduce disease. For example, in a process known as the dilution effect[4], the risk of disease is lowered if the most favourable host for a pathogen is a small component of a diverse host community. Alternatively, pathogens may grow rapidly in certain hosts, and the resulting high numbers of diseased individuals might allow the pathogen to 'spill over' and contribute to increased infection of other hosts[5]. Although these topics are directly relevant to human and wildlife diseases, they are hard to study with mobile hosts. By contrast, Parker et al. show that such questions can be examined directly with plants.

The authors' field study combined observational and experimental approaches. The team first documented the average amount of

diseased leaf tissue on 43 plant species (Fig. 1). Rarer species had lower disease levels, as expected given the typical positive association between host abundance and disease. The authors also determined the evolutionary relationships of the 43 species and found that species that had fewer relatives had lower disease levels. Next, they used a model based on an independent data set of 210 plant genera and 212 fungal pathogens to explore evolutionary relationships between host plants and pathogens. This model, combined with knowledge of host abundance, provided the best predictions of disease levels in the field.

Finally, the team introduced 44 other plant species to the field sites. These were chosen for their evolutionary diversity: some were closely related to species at the field sites, whereas others were distantly related. The authors were able to successfully predict disease levels on these new species using their model of host–pathogen evolutionary relationships. In a nutshell, the message of this paper is that disease levels depend not only on ecological factors (host abundance) but also on the evolutionary relationships of the host community.

These results are perhaps not surprising — after all, many diseases are specific to plants of the same genus or family. However, Parker and colleagues' study has three main strengths: the authors quantified the relationship between disease levels and host evolutionary relationships in the context of natural disease transmission; they made explicit predictions and tested them experimentally; and the general nature of their model of host–pathogen evolutionary relationships means that it can be widely used by other researchers.

This work contributes to ongoing discussions about the diversity and composition of communities. For example, the authors' field sites were 'overdispersed', meaning that they contained fewer related plant species than would be expected by chance given the possible species in the region. This pattern has been thought to be a consequence of resource competition, because unrelated species may compete less with each other[6]. Parker and colleagues' findings suggest an alternative mechanism: that pathogens shape community structure if closely related host species are less

Figure 1 | Health amid disease. Parker et al.[3] find that fungal infections are more likely to spread between closely related species, such as the grasses shown here in their field site, than to more distantly related species, such as this wild radish.

successful owing to increased disease levels.

The study also has applied significance. For instance, it implies that less disease should occur in plant communities with combinations of species from distinct evolutionary lineages. This information is relevant to designs for intercropping (when two or more crops are grown in close proximity) and forest-tree mixtures[7]. Furthermore, the findings relate to the major challenge of predicting the fate of newly introduced species. Given a list of resident host species, one could use the authors' approach to explore whether a certain introduced species is likely to share pathogens with others, which may affect its invasion success. The abundances of individual host species (both native and non-native) will, of course, also contribute to determining disease levels.

Although Parker et al. focused on leaf-infecting fungi, evolutionary relationships are important for characterizing the host range for diverse natural enemies of plants,

including other microbes and herbivores[8]. Future work should include mechanistic studies (for example, to assess whether there are certain plant traits that drive relationships between pathogens and host evolutionary history) and studies that explore the generality of these field results. For example, the nonlinear relationship in Figure 3 of the paper[3] suggests that recent evolutionary history is particularly relevant to predicting disease levels. Would this type of relationship be expected across other plant communities and natural enemies? Future research should also explore the roles of other ecological factors in disease prediction, such as the mode and distance of pathogen dispersal, aggregation patterns of plant species and the spatial scale of studies. Finally, it is worth noting that the ecological interpretations of these studies assume an inverse relationship between disease expression and host fitness — but it is challenging to perform experiments that assess disease effects on fitness, especially in diverse field communities with many plants and pathogens.

Parker and colleagues' careful research is the latest in a suite of studies that emphasize the importance of evolutionary history in ecological processes[6,7,9]. This line of enquiry illustrates the continued relevance of a famous quote from biologist Theodosius Dobzhansky[10]: "Nothing in biology makes sense except in the light of evolution." ∎

Helen M. Alexander *is in the Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas 66045, USA. e-mail: halexander@ku.edu*

1. Mordecai, E. A. *Ecol. Monogr.* **81**, 429–441 (2011).
2. Maron, J. L., Marler, M., Klironomos, J. N. & Cleveland, C. C. *Ecol. Lett.* **14**, 36–41 (2011).
3. Parker, I. M. *et al. Nature* **520**, 542–544 (2015).
4. Keesing, F., Holt, R. D. & Ostfeld, R. S. *Ecol. Lett.* **9**, 485–498 (2006).
5. Power, A. G. & Mitchell, C. E. *Am. Nat.* **164**, S79–S89 (2004).
6. Webb, C. O., Ackerly, D. D., McPeek, M. A. & Donoghue, M. J. *Annu. Rev. Ecol. Syst.* **33**, 475–505 (2002).
7. Castagneyrol, B., Jactel, H., Vacher, C., Brockerhoff, E. G. & Koricheva, J. *J. Appl. Ecol.* **51**, 134–141 (2014).
8. Gilbert, G. S., Magarey, R., Suiter, K. & Webb, C. O. *Evol. Appl.* **5**, 869–878 (2012).
9. Gilbert, G. S. & Webb, C. O. *Proc. Natl Acad. Sci. USA* **104**, 4979–4983 (2007).
10. Dobzhansky, T. *Am. Biol. Teach.* **35**, 125–129 (1973).

nature INSIGHT

ORIGIN AND EVOLUTION OF VERTEBRATES

nature INSIGHT

ORIGIN AND EVOLUTION OF VERTEBRATES

Cover illustration
Jasiek Krzysztofiak

To celebrate the golden jubilee of *On the Origin of Species*, in 1909, the Linnean Society of London held a special meeting on a hot biological topic of the day — the origin of the vertebrates. Such was the lack of consensus that one commentator, the zoologist T. R. R. Stebbing, wrote that "the disputants agreed on one single point, namely, that their opponents were all in the wrong."

The problem is easily stated — vertebrates have so many special features, from large brains to complex physiologies to unique tissues such as enamel and bone — that their evolution from invertebrates is obscure. The question had intrigued Aristotle, and foxed minds as keen as those of William Bateson and Thomas Hunt Morgan, who, by way of finding a more rewarding problem, went off to discover genetics instead.

The same tools that Bateson and Hunt Morgan helped to create have now returned to address the old problem. Although our understanding is far from complete, it is much better than it was even 20 years ago, and is summarized in this collection of reviews.

Nicholas Holland and colleagues set out how the varied theories advanced to explain vertebrate origins, before Lowe *et al.* show how they fit in to the deuterostomes, a larger branch of the animal kingdom. Diogo *et al.* add new perspectives to a central question of vertebrate origins, namely, the origin of the head. Marianne Bronner and colleagues then look at the embryonic tissue known as neural crest, another uniquely vertebrate feature. Philippe Janvier surveys the wealth of newly found, and often curious, fossil evidence, and Martin Brazeau and Matt Friedman chart the evolution of jawed vertebrates from jawless forms. If Stebbing was able to peruse this collection, I hope he would agree that we have come a long way.

Henry Gee
Senior Editor

CONTENTS

Scenarios for the making of vertebrates

Nicholas D. Holland[1], Linda Z. Holland[1] & Peter W. H. Holland[2]

Over the past 200 years, almost every invertebrate phylum has been proposed as a starting point for evolving vertebrates. Most of these scenarios are outdated, but several are still seriously considered. The short-range transition from ancestral invertebrate chordates (similar to amphioxus and tunicates) to vertebrates is well accepted. However, longer-range transitions leading up to the invertebrate chordates themselves are more controversial. Opinion is divided between the annelid and the enteropneust scenarios, predicting, respectively, a complex or a simple ancestor for bilaterian animals. Deciding between these ideas will be facilitated by further comparative studies of multicellular animals, including enigmatic taxa such as xenacoelomorphs.

Biologists have considered nearly every major taxon of animals as the key starting point for the evolution of vertebrates. We survey these ideas, many of which are no longer tenable in the light of subsequent advances in biology, and then concentrate on the few scenarios that are currently the subject of major research programmes. Lamarck was the first to propose an evolutionary conversion from an invertebrate to a vertebrate. In 1809, he depicted a phylogenetic tree, including an invertebrate-to-vertebrate transition in which molluscs gave rise to fishes[1]. During the next few decades, several others speculated on how body plans of invertebrates and vertebrates might be related; however, those biologists were generally in search of an underlying unity of organismal design. Evolution was not explicitly mentioned by key figures such as Geoffroy Saint-Hilaire[2], although one senses that he was on the verge of believing in it.

Aside from Lamarck's proposal, explicitly evolutionary schemes that derived vertebrates from invertebrates started appearing only after the publication of *On the Origin of Species* in 1859. In general, the scenarios were based on the morphology of developmental stages and adults of extant animals. Palaeontological evidence was considered less often[3–5], and molecular evidence was not widely considered until the 1980s with the advent of molecular phylogenetics and evolutionary developmental biology.

In Fig. 1, scenarios for the origin of vertebrates are arranged on a timeline extending from the publication of *On the Origin of Species* to the present. The references are broadly divided into those focused on larval type and those concerned with adults — a dichotomy reflecting two opposing views of life-history evolution. The first considers pelagic larvae as primal with benthic stages added later, and the second considers benthic stages as primitive with pelagic larvae interpolated later. Classification of the scenarios is not straightforward because relatively few proposed a linear ancestor–descendant relationship. More commonly, they were presented in the context of branching, sister-group relationships. For example, if enteropneusts were considered the sister group of the chordates (as in Fig. 2a), the ancestral node is often referred to as enteropneust-like. Finally, when a given scheme involves an evolutionary pathway through several major taxa to the vertebrates, the scenario is named for the invertebrate group receiving the most attention from the original author.

Scenarios currently the subject of active research

Contemporary research on the origin of vertebrates from invertebrates falls into two broad categories: the short-range transition from invertebrate chordates (amphioxus-like and tunicate-like ancestors) to vertebrates, and longer-range transitions from the base of bilaterally symmetrical animals or from the base of deuterostomes to vertebrates. Scenarios starting with invertebrate chordates are less controversial than the two long-range scenarios being actively studied: the annelid and the enteropneust theory.

Invertebrate chordate to vertebrate transition

The nearest relatives of the vertebrates are the invertebrate chordates, although it is still not settled whether chordate evolution should be considered from the viewpoint of larvae being primal[6] or larvae being interpolations[7,8]. Although invertebrate chordate scenarios ignore the deeper history of the vertebrate lineage, they still centre on events initiated more than 500 million years ago and involve remarkable evolutionary changes that are considered in companion reviews in this issue. Recently, the major chordate taxa were rearranged (Fig. 2b) on the basis of morphology and molecular phylogenetics, which have decisively shown that amphioxus is the sister group to tunicates and vertebrates[9,10]. The new arrangement implies that the tunicates have secondarily lost segmentation, coeloms and kidneys, but are vertebrate-like in features such as intercellular tight junctions, proto-neural crest, striated heart muscles, proto-placode derivatives and voluminous blood plasma with abundant circulating corpuscles.

The annelid theory

The first of the two long-range scenarios is the annelid theory. When initially published 140 years ago, it proposed a direct conversion of annelid worms into vertebrates[11,12]. Now, however, the starting point is often considered to be an annelid-like urbilaterian[13] (Fig. 2a, b). The annelid theory has its roots in arthropod biology, because these two groups were long considered to be very close relatives (Fig. 2a), and results for one were generally considered to be valid for the other.

In the original annelid scenario, Dohrn[11] started with a worm that inverted the body on the way to evolving into a vertebrate, thus positioning the old mouth on the top of the head and necessitating the formation of a new mouth on the ventral side of the body; thereafter, the old mouth disappeared, while the new one persisted (Fig. 3a–c). Several of Dohrn's colleagues modified his scenario in attempts to improve it[14–21], but the theory went into eclipse early in the twentieth century when the bilaterian animals were rearranged into two superphyla — the protostomes and the

[1]Marine Biology Research Division, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California 92093, USA. [2]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.
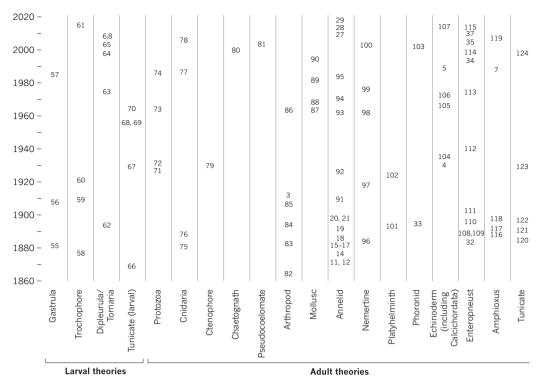
Figure 1 | Scenarios for the invertebrate-to-vertebrate transition. Each scenario is categorized according to the larval type or to the taxon of adult invertebrate proposed as ancestral to the vertebrates[3–8,11,12,14–21,27–29,32–35,37,55–124]. For prolix authors, only their most inclusive publications are given. Also omitted are references (typically textbooks) that repeat previous ideas without adding new information. References to problematic fossil ancestors of vertebrates are not included (except calcichordates, which are considered to be echinoderms here).

Larval theories: Gastrula, Trochophore, Dipleurula/Tornaria, Tunicate (larval)

Adult theories: Protozoa, Cnidaria, Ctenophore, Chaetognath, Pseudocoelomate, Arthropod, Mollusc, Annelid, Nemertine, Platyhelminth, Phoronid, Echinoderm (including Calcichordata), Enteropneust, Amphioxus, Tunicate

deuterostomes. The resulting relocation of annelids and arthropods at a considerable phylogenetic distance from the vertebrates (Fig. 2a) weakened the idea of a complex urbilaterian and shifted opinion towards a simple urbilaterian, which was imagined to be rather like an acoel flatworm that independently gave rise to annelids and vertebrates with their complex, but only superficially similar, body plans.

In the 1990s, advances in developmental genetics — again with arthropods leading the way — set the stage for the revival of the annelid theory. The fly *dpp* gene was found to be expressed dorsally and to have dorsalizing activity, whereas the homologous frog *bmp4* was expressed ventrally and found to have ventralizing activity[22]. Arendt and Nübler-Jung interpreted this pattern as support for homology between arthropod and vertebrate nerve cords and indicative of a dorsoventral inversion of the body during the invertebrate-to-vertebrate transition[23]. The proposed nerve-cord homology was strengthened by the discovery that the fly *sog* gene was expressed ventrally and had ventralizing activity, whereas the homologous frog *chordin* gene was expressed dorsally and had dorsalizing activity. In addition, *sog/chordin* and *dpp/bmp4* antagonized one another to establish a dorsoventral axis that was reversed between flies and frogs[24]. Additional support came from the finding that neural progenitor cells in the central nervous system (CNS) were organized in longitudinal bands each characterized by a distinctive suite of gene expression that was homologous between flies and vertebrates, and that gene expression in these bands was comparable mediolaterally in both organisms[25].

The developmental genetic comparison between arthropods and vertebrates[22–25], reinforced by details from neurochemistry and neural circuitry, favoured the revival of the inverted annelid theory. Direct comparisons between annelids and vertebrates also revealed commonalities in anterior–posterior regionalization by Hox genes[26], genetic specification of several kinds of nerve cells[27,28] and the formation of notochord-like structures[29]. As already mentioned, the revived annelid scenario posits the evolution of an already complex urbilaterian ancestor into a vertebrate. Such a transition would be most parsimonious if it proceeded through consistently complex intermediates. However, some features, such as segmentation and a clearly centralized nerve cord, are absent from several taxa associated with the presumed evolutionary lineage that leads to the vertebrates, possibly due to secondary losses. Such losses would have occurred in echinoderms, at least some hemichordates and
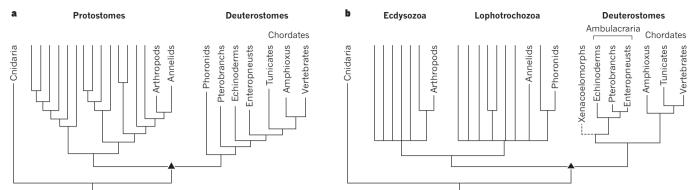
xenacoelomorphs, although the deuterostome nature of the last has not yet been firmly established[30,31]. Continuity between annelid-like ancestors and vertebrates could be strengthened if complex, segmented fossils of basal deuterostomes were known. Although several such fossils have been proposed as ancestral deuterostomes (vetulicolians and *Herpetogaster*), their taxonomic affinities remain highly controversial.

The enteropneust theory

The second long-range scenario of vertebrate origins currently under active study is the enteropneust theory. These marine worms (Fig. 3d), characterized by three body regions (proboscis, collar and trunk), belong to the Hemichordata, a phylum that also includes the minute ptero-branchs (Fig. 3e), which comprise a flattened oral shield corresponding to the enteropneust proboscis, a collar extending into tentacle-fringed arms and a trunk. According to the original enteropneust theory[32], proposed by Bateson in 1886, the body axis of enteropneusts was not inverted relative to that of vertebrates. For him, the stomochord (Fig. 3d) corresponded to a vertebrate notochord, the collar cord (which he considered dorsal) corresponded to the vertebrate CNS, and the pharyngeal gill slits in both groups were homologous. Such an enteropneust was much like a vertebrate except that it lacked segmented musculature along the anterior–posterior axis. At the time, Bateson was uncertain about the deeper evolutionary source of the enteropneusts, although he tentatively suggested that they might have evolved from nemerteans or even tunicates. However, at the close of the nineteenth century, Masterman[33] proposed what seemed to be a firmer connection between enteropneusts and the rest of the animal kingdom through relatively complex precursors — pterobranchs (already mentioned) and the worm-like phoronids, which live mostly buried, but extend their tentacle crown into the sea water.

Through much of the twentieth century, Bateson's hypothesis, although not universally accepted, persisted. This inactivity ended in 1996, when Nübler-Jung and Arendt made a striking alteration[34]. They proposed that enteropneusts had an annelid-like CNS comprising three contiguous nerve tracts (the collar cord, the circumenteric nerve ring and the trunk ventral nerve cord), all recognizable by their giant nerve fibres. Such an enteropneust (Fig. 3f) complemented their earlier revival of the annelid theory[23] by approximating an intermediate stage in the conversion of a complex urbilaterian into a vertebrate. Because this CNS was oriented

Figure 2 | Simplified trees of metazoan animal life. Taxa not mentioned in this Review are indicated by unlabelled branches (that are reduced in number and intended to be diagrammatic); the position of the Urbilateria is indicated by a triangle. **a**, Morphology-based tree[65]. **b**, Sequence-based tree[125]; the dashed line emphasizes the current uncertainty about the placement of the xenacoelomorphs.

as in annelids, the conversion into a vertebrate-like descendant (Fig. 3g) would require dorsoventral inversion, in contrast to Bateson's original scenario. While Nübler-Jung and Arendt were revising the enteropneust theory[34], molecular phylogenetics revealed that the relatively complex phoronids are neither deuterostomes nor their close relatives[30] (Fig. 2b). One interpretation of the new phylogeny was that the ancestors of the enteropneusts had relatively simple body plans — traceable back to an even simpler urbilaterian. The new phylogenetic arrangement triggered the definitive revival of the enteropneust theory that is still in progress.

The chief proponent of this newest revival of the enteropneust theory is Lowe, who gathered support for it with developmental genetic studies (see Review on page 456). He first considered a score of genes with homologues patterning the vertebrate CNS along its anterior–posterior axis[35]. Most of these enteropneust genes were expressed in the same anterior–posterior order as their homologues in the vertebrate CNS — but in annular bands of ectoderm and not in any tissue that might be interpreted as a CNS[35]. He concluded that the nervous system lacked any CNS component and consisted exclusively of an ectodermal nerve net. Although vertebrate homologues of many of the genes studied by Lowe help to establish borders separating neuronal populations in the vertebrate CNS[36], no corresponding neuroanatomical or neurophysiological discontinuities have yet been found in any enteropneust tissue. In Lowe's original scenario, the transition of enteropneust-like ancestors into vertebrates involved a loss of most of the ectodermal neurons, except along the midline of the body, where a CNS was elaborated. By similar, but independent paths, the dispersed nerve net of a structurally simple urbilaterian would have given rise to the complex CNS of annelids and arthropods.

Lowe subsequently studied the genes involved in establishing the dorsoventral axis of enteropneusts[37] and found that *BMP* and *chordin* were expressed, respectively, on the dorsal and ventral sides — if the body is assumed to be oriented similarly to annelids and arthropods. However, unlike the situation in amphioxus and vertebrates, upregulation experiments failed to alter neuron distribution, although some non-neural structures (the mouth, for example) were repositioned as expected. These results suggested that the *BMP*–*chordin* axis initially patterned exclusively non-neural structures and only later in evolution became linked to positioning neurons. This linkage to neural development was thought to have occurred independently in annelids, arthropods and vertebrates. Lowe[37] considered, but initially rejected, the converse possibility: that the relation between dorsoventral signalling and nervous-system development was ancient and was secondarily lost in the lineage leading to the enteropneusts.

More recent work challenges one point in the revived enteropneust scenario: that no CNS is present. First, Nomaksteinsky *et al.*[38] suggested that the proboscis plexus, collar cord, circumenteric nerves, and trunk dorsal and ventral cords have some properties of a CNS — cell bodies of neurons are present and extend their neurites into an adjacent neuropil — and that the epidermis outside the nerve cords includes only widely scattered nerve cells representing a sparse peripheral nervous system instead of a nerve net. In addition, Cunningham and Casey[39] found enteropneust neuronal

marker genes expressed along both the dorsal and ventral cord of the trunk, which they too suggested might be parts of a CNS. Neither study could resolve the dorsoventral orientation of the enteropneust body. In an attempt to answer this question, the left–right asymmetry of *Nodal* gene expression was compared during development of several deuterostomes. Right-sided expression in echinoderms and enteropneusts contrasted with left-sided expression in vertebrates, indicating that the dorsoventral axis of vertebrates is indeed inverted relative to that of echinoderms and enteropneusts[40,41]. As a caveat, however, although *Nodal* is involved in establishing the left–right axes of echinoderms and vertebrates, it evidently has no comparable functional role in enteropneusts[42].

To complicate matters further, Miyamoto and Wada[43] found that the endoderm of the enteropneust stomochord and the roof of the buccal cavity are sources of Hedgehog signals that evidently induce and pattern the collar nerve cord. This parallels Hedgehog signalling from the notochord to the nascent neural tube during vertebrate development. Their data could be interpreted to mean that dorsoventral inversion did not take place during the enteropneust-to-vertebrate transition, that the stomochord is homologous to a notochord, and that the collar cord corresponds to at least part of the vertebrate CNS. These conclusions are close to those reached by Bateson in his original scenario[32], although Miyamoto and Wada acknowledge that co-option of gene networks cannot be ruled out. These disagreements about the enteropneust nervous system seem likely to be resolved by additional neuroanatomical studies. However, that would still leave the nature of the urbilaterian unsettled, which will be considered in the next section.

Progress, problems and prospects

At the end of an argumentative symposium on the origin of vertebrates a century ago[44], one participant summed up progress with the mischievous words: "When we return home and our friends gleefully enquire, 'What then has been decided as to the Origin of Vertebrates?', so far we seem to have no reply ready, except that the disputants agreed on one single point, namely, that their opponents were all in the wrong." Although prospects for solving the riddle of vertebrate origins at that time did not look good, there has been progress. In particular, we now know where vertebrates fit in the animal phylogenetic tree. This knowledge helps to refine the remaining questions. To start with, we can consider an evolutionary tree as including a nested series of ancestors, each defining a different node of the tree, progressively deeper in time. As we climb down the tree, back in time from the living vertebrates, we encounter each ancestral node in turn. As we proceed, we should not be asking what did the ancestor of vertebrates look like? But instead what did each successive ancestor of the vertebrates look like? This logic can be applied to the node-based ancestors, but we should remember that there must have been an unbroken, genealogically connected series of ancestors between each node that are all but invisible to comparative biology based on living taxa.

Logically, the most recent node-based ancestor of all living vertebrates was itself a vertebrate, and possessed characters shared by lampreys,
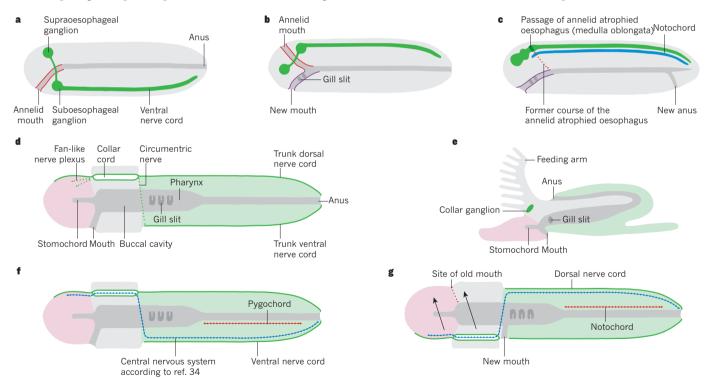
hagfish and jawed vertebrates. This animal, living more than half a billion years ago, had a well-developed head and brain, complex cranial sense organs, segmented musculature and a vertebral column (recently shown to be present but secondarily reduced in hagfish[45]), but no jaws or paired fins. The ancestor also probably shared the genome duplications that set vertebrates apart from other deuterostomes. The subsequent course of evolution in the vertebrates is considered in several companion Reviews in this Insight; however, here we are concerned with looking the other way — towards the invertebrate roots of the vertebrates.

The two closest lineages to the vertebrates are the tunicates and the cephalochordates (such as amphioxus). Like vertebrates, both are chordates. The chordate ancestor had segmented muscle blocks, a notochord and a dorsal CNS. It also probably gathered food particles on secretions produced by a glandular endostyle located in an expanded, perforated pharynx. Controversy remains over what the head region of this long-extinct ancestor looked like, because the anterior region of tunicates (or their larvae) is so different from that of amphioxus. Did this ancestor have mesodermal somites (segments) in its anterior region, like a modern amphioxus, or was the anterior unsegmented as it is in tunicate larvae? This may sound like a minor issue, but it is important to resolve if we wish to understand how our own head and brain arose in evolution. This old debate remains unsettled, and more work is needed to compare gene expression and cellular fates in the cranial regions of each chordate group as well as between the cranial and somitic mesoderm of vertebrates.

At the next node-based ancestor, the basal deuterostome, the rival claims of the annelid and enteropneust theories first begin competing for our attention. The chordates are the sister group to the Ambulacraria[46], a clade comprising enteropneusts, pterobranchs and echinoderms (Fig. 2b).

Somewhere in the mix may also be the acoels and nemertodermatids (tiny animals with an inconspicuous nervous system) and possibly the larger, but similarly simple, xenoturbellids (here, we will accept the unification of these three groups as xenocoelomorphs[31,47]). The placement of the xenacoelomorphs in the evolutionary tree is also debated; for example, molecular phylogenies that place them as sisters to Ambulacraria — plus or minus the chordates — do not sit easily with other features such as their simple Hox gene cluster[47]. To understand the importance of xenocoelomorphs, we need to consider the common ancestor of Ambulacraria and Chordata. This animal in our series of vertebrate ancestors possessed pharyngeal slits (homologous in enteropneusts and chordates[48]), but what else? Did it have a brain and a CNS, for example? Chordates have a dorsal centralized nerve cord, whereas at least echinoderms have a dispersed nervous system that may be relatively condensed in some regions and not generally considered a CNS, although there is an element of subjectivity in deciding what constitutes a CNS. The putative CNS nature of enteropneust nerve cords[38] has been noted earlier. A similar debate surrounds xenacoelomorphs: xenoturbellids are not considered to have a brain, whereas acoels and nemertodermatids have small anterior aggregations of neural tissue that some have considered to be brain-like[47]. If xenacoelomorphs are basal in the deuterostomes, one might envisage the common ancestor of chordates and ambulacrarians to be enteropneust-like in lacking a clear CNS and a 'brain', although secondary simplification might have occurred[31].

Does this mean that the enteropneust theory wins over the annelid theory? Unfortunately, things are not simple. First, centralized nerve cords are widely distributed (although far from the rule) among bilaterian animals. Thus, concluding that the urbilaterian (and in turn the later ambulacrarian and chordate common ancestor) possessed a CNS would not be

Figure 3 | Annelid and enteropneust theories. a, An annelid with a central nervous system (CNS; green) comprising supraoesophageal and suboesophageal ganglia, circumoesophageal connectives and ventral nerve cord. **b,** Dorsoventral inversion[11] produces a new foregut (purple) penetrated by gill slits. **c,** Annelid-to-vertebrate transition. The new foregut persists, but the old one atrophies, permitting union of the supra- and suboesophageal ganglia into a vertebrate-like brain. A notochord (blue) originates from connective tissue surrounding the nerve cord, and a new anus opens. **d,** Enteropneust according to Bateson[32], showing proboscis (pink), collar (grey) and trunk (light green). The ventral mouth opens into a buccal cavity, giving off a small diverticulum (the stomochord) anteriorly and connecting with the pharynx posteriorly. Gill slits penetrate either side of the pharynx,

and the post-pharyngeal gut ends posteriorly at the anus. **e,** A pterobranch hemichordate (*Rhabdopleura*), comprising a cephalic shield (pink), collar with feeding arms (grey) and trunk (light green). **f,** Enteropneust as conceived by Nübler-Jung and Arendt[34] with the blue line showing the extent of the CNS. The red line indicates the pygochord. **g,** Proposed inversion during enteropneust-to-vertebrate transition[34]. The pygochord becomes the notochord; the trunk ventral nerve cord becomes the dorsal nerve cord; a dorsal shift of the proboscis plexus and collar cord (arrows) supplies anterior brain regions; and a new mouth forms, while the old one disappears. The transition proposed by Nübler-Jung and Arendt (shown here between **f** and **g**) has now been supplanted by the more current scenario of Lowe[35,37], which is covered in detail by the Review on page 456.

unreasonable. This would imply secondary reduction in basal deuterostomes. A key issue is whether the deuterostome ancestor was segmented along the body axis[49]. This question is inextricably linked to the question of whether the urbilaterian was also segmented. Several distantly related invertebrates are segmented along the body axis, including arthropods and annelids. If their segmentation is homologous with that of chordates, then, as has been suggested[50], the ancestors of both deuterostomes and protostomes were segmented, and enteropneusts lost their segments.

However, the segmentation issue is still vigorously debated. Molecular similarities in the control of segmentation between arthropods and some annelids are striking[6,51], and there are commonalities in gene expression between mesodermal segmentation in these two phyla and in chordates[52]. Even so, deciding whether the similarities in segmentation are due to inheritance from a common ancestor or to independent co-option of parts of the same molecular machinery[53] is not straightforward. New modes of segmentation (in the broad sense[49]), such as hindbrain rhombomeres of vertebrates and reiterated pharyngeal slits of deuterostomes, can arise in evolution. Molecular and cellular studies of segmental patterning mechanisms across the animal kingdom and the nature of cycling gene networks are needed to tackle this issue. At present, therefore, we suggest that the common ancestor of ambulacrarians and chordates probably mixed the enteropneust character of pharyngeal slits and the annelid and chordate character of a centralized nerve cord. We cannot say with certainty that this ancestor was segmented along the body axis. Finally, some salient chordate characters seem to be novelties without precedents in either annelids or enteropneusts. For example, the organization of vertebrate muscle blocks working together with a notochord for active undulatory swimming[7]; this arrangement differs distinctively from the disposition of circular and longitudinal muscles in most other bilaterian animals.

If we now consider the next deepest node-based ancestor of the vertebrates, the urbilaterian ancestor to all bilateral animals, we can use the same logic as earlier, and many of the same data, to approach the reconstruction of the body plan. To pick up on just the three key morphological features discussed earlier — pharyngeal slits, a central nerve cord and segments — we deduce that it lacked pharyngeal slits and might have possessed a central nerve cord. However, there is too much uncertainty to decide whether it had segments along the body axis. It is reasonable to assume that the urbilaterian was unlike any animal alive today, but shared characters both with modern annelids and with modern enteropneusts. To turn this around, each of these two living groups seem to retain some of the characters from their, and our, distant ancestor.

We conclude, therefore, that the annelid and enteropneust scenarios are both partly correct. Some of the early proponents of the vast range of scenarios for the origin of vertebrates (Fig. 1) viewed living animals as proxies for long extinct ancestors. In reality, more progress has been made by comparing living animals with one another to deduce the combinations of morphological characters present in ancestors, a task that requires critical evaluation of homology, incorporating developmental, cellular and molecular approaches in an ever-widening range of animal taxa. Ultimately, a wealth of reliable and detailed information over a wide spectrum of taxa will be needed to sort out relationships among the animal phyla and their component characters[53,54]. This Review began with an appreciation of the older ideas in the field, and some of these will continue to guide us as we move ahead with technological advances and new discoveries in biology and palaeontology to gain insights into the origin of the vertebrates and our own distant history. ■

1. Lamarck, J. B. *Philosophie Zoologique* [in French] Vol. 2 (Dentu, 1809).
   The first explicitly evolutionary derivation of vertebrates from invertebrates.
2. Geoffroy St-Hilaire, E. Considérations générales sur la vertèbre [in French]. *Mém. Mus. Hist. Nat.* **9,** 89–119 (1822).
3. Patten, W. *The Evolution of the Vertebrates and Their Kin* (Blakiston's Son, 1912).
4. Gislén, T. Affinities between the Echinodermata, Enteropneusta, and Chordania. *Zool. Bidr. Uppsala* **12,** 199–304 (1930).
5. Jefferies, R. P. S. *The Ancestry of the Vertebrates* (British Museum Natural History, 1986).
   This scholarly work, even if no longer accepted in its broad outline, was centrally important for reawakening interest in the invertebrate-to-vertebrate transition
6. Satoh, N. An aboral-dorsalization hypothesis for chordate origin. *Genesis* **46,** 614–622 (2008).
7. Gans, C. & Northcut, R. G. Neural crest and the origin of the vertebrates: a new head. *Science* **220,** 268–274 (1983).
   This influential scenario did much to focus the attention of vertebrate biologists back on questions about the evolutionary origin of their group.
8. Northcutt, R. G. The new head hypothesis revisited. *J. Exp. Zool.* **304B,** 274–297 (2005).
9. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439,** 965–968 (2006).
10. Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453,** 1064–1071 (2008).
11. Dohrn, A. *Der Ursprung der Wirbelthiere und das Princip des Functionswechsels: genealogische Skizzen* [in German] (Engelmann, 1875).
    This is the original annelid theory.
12. Semper, C. Die Stammesverwandtschaft der Wirbelthiere und Wirbellosen [in German]. *Arb. Zool.-Zootom. Inst. Würzburg* **2,** 25–76 (1875).
13. Balavoine, G. & Adoutte, A. The segmented urbilateria: a testable scenario. *Integr. Comp. Biol.* **43,** 137–147 (2003).
14. Eisig, H. Der Nebendarm der Capitelliden und seine Homologa [in German]. *Zool. Anz.* **1,** 148–152 (1878).
15. Kleinenberg, N. Die Entstehung des Annelids aus der Larve von Lopadorhynchus, nebst Bemerkungen über die Entwicklung anderer Polychaeten [in German]. *Z. Wiss. Zool.* **44,** 1–227 (1886).
16. Van Beneden, E. & Julin, C. Recherches sur la morphologie des tuniciers [in French]. *Arch. Biol.* (Liege) **6,** 237–476 (1886).
17. Koehler, R. Sur la parenté du Balanoglossus [in French]. *Zool. Anz.* **9,** 506–507 (1886).
18. Beard, J. Some annelidan affinities in the ontogeny of the vertebrate nervous system. *Nature* **39,** 259–261 (1889).
19. Kennel, J. Ueber die Ableitung der Vertebratenaugen von den Augen der Anneliden [in German]. *Sitzungsber. Naturforsch. Ges. Univ. Dorpat.* **9,** 408–411 (1892).
20. Minot, C. S. Cephalic homologies. A contribution to the determination of the ancestry of the vertebrates. *Am. Nat.* **31,** 927–943 (1897).
21. Bernard, H. M. A new reading for the annulate ancestry of the Vertebrata. *Nat. Sci.* **13,** 17–30 (1898).
22. Jones, C. M., Lyons, K. M., Lapan, P. M., Wright, C. V. E. & Hogan, B. L. M. DVR-4 (bone morphogenetic protein-4) as a posteriorventralizing factor in *Xenopus* mesoderm induction. *Development* **115,** 639–647 (1992).
23. Arendt, D. & Nübler-Jung, K. Inversion of dorsoventral axis? *Nature* **371,** 26 (1994).
    This publication launched the current revival of the annelid theory.
24. Holley, S. A. *et al.* A conserved system for dorso-ventral patterning in insects and vertebrates involving *sog* and *chordin*. *Nature* **376,** 249–253 (1995).
    This application of developmental genetic data to classic questions of animal phylogeny attracted much attention to the young field of evolutionary developmental biology.
25. Arendt, D. & Nübler-Jung, K. Comparison of early nerve cord development in insects and vertebrates. *Development* **126,** 2309–2325 (1999).
26. Kulakova, M. *et al.* Hox gene expression in larval development of the polychaetes *Nereis virens* and *Platynereis dumerilii* (Annelida, Lophotrochozoa). *Dev. Genes Evol.* **217,** 39–54 (2007).
27. Denes, A. S. *et al.* Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in Bilateria. *Cell* **129,** 277–288 (2007).
28. Tomer, R., Denes, A. S., Tessmar-Raible, K. & Arendt, D. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell* **142,** 800–809 (2010).
29. Lauri, A. *et al.* Development of the annelid axochord: insights into notochord evolution. *Science* **345,** 1365–1368 (2014).
30. Halanych, K. M. The new view of animal phylogeny. *Annu. Rev. Ecol. Evol. Syst.* **35,** 229–256 (2004).
31. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* **470,** 255–260 (2011).
32. Bateson, W. The ancestry of the Chordata. *Q. J. Microsc. Sci.* **26,** 535–571 (1886).
    This article is the original enteropneust theory.
33. Masterman, A. T. On the Diplochorda. I. The structure of Actinotrocha. *Q. J. Microsc. Sci.* **40,** 281–338 (1897).
34. Nübler-Jung, K. & Arendt, D. Enteropneusts and chordate evolution. *Curr. Biol.* **6,** 352–353 (1996).
35. Lowe, C. J. *et al.* Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* **113,** 853–865 (2003).
    This publication launched the current revival of the enteropneust theory.
36. Puelles, L. & Ferran, J. L. Concept of neural genoarchitecture and its genomic fundament. *Front. Neuroanat.* **6,** 47 (2012).
37. Lowe, C. J. Molecular genetic insights into deuterostome evolution from the direct-developing hemichordate *Saccoglossus kowalevskii*. *Phil. Trans. R. Soc. B* **363,** 1569–1578 (2008).
38. Nomaksteinsky, M. *et al.* Centralization of the deuterostome nervous system predates chordates. *Curr. Biol.* **19,** 1264–1269 (2009).
39. Cunningham, D. & Casey, E. S. Spatiotemporal development of the embryonic nervous system of *Saccoglossus kowalevskii*. *Dev. Biol.* **386,** 252–263 (2014).
40. Duboc, V., Röttinger, E., Lapraz, F., Besnardeau, L. & Lepage, T. Left-right asymmetry in the sea urchin embryo is regulated by nodal signaling on the right side. *Dev. Cell* **9,** 147–158 (2005).
41. Wlizla, M. Evolution of Nodal signaling in Deuterostomes: insights from *Saccoglossus kowalevskii*. PhD thesis, Univ. Chicago (2011).
42. Röttinger, E., Duboc, T. & Martindale, M. Q. Investigating the role of the Nodal signaling pathway in a indirect developing hemichordate, *Ptychodera flava*. *Integr. Comp. Biol.* **50,** abstract, e144 (2010).

43. Miyamoto, N. & Wada, H. Hemichordate neurulation and the origin of the neural tube. *Nature Commun.* **4,** 2713 (2013).
44. Gaskell, W. H. *et al.* Origin of vertebrates. *Proc. Linn. Soc. Lond.* **122,** 9–50 (1910).
45. Ota, K. G., Fujimoto, S., Oisi, Y. & Kuratani, S. Identification of vertebra-like elements and their possible differentiation from sclerotomes in the hagfish. *Nature Commun.* **2,** 373 (2011).
46. Furlong, R. F. & Holland, P. W. H. Bayesian phylogenetic analysis supports monophyly of ambulacraria and of cyclostomes. *Zoolog. Sci.* **19,** 593–599 (2002).
47. Achatz, J. G., Chiodin, M., Salvenmoser, W., Tyler, S. & Martinez, P. The Acoela: on their kind and kinships, especially with nemertodermatids and xenoturbellids (Bilateria incertae sedis). *Org. Divers. Evol.* **13,** 267–286 (2013).
48. Ogasawara, M., Wada, H., Peters, H. & Satoh, N. Developmental expression of *Pax1/9* genes in urochordate and hemichordate gills: insight into function and evolution of the pharyngeal epithelium. *Development* **126,** 2539–2550 (1999).
49. Graham, A., Butts, T., Lumsden, A. & Kiecker, C. What can vertebrates tell us about segmentation? *EvoDevo* **5,** 24 (2014).
50. De Robertis, E. M. The molecular ancestry of segmentation mechanisms. *Proc. Natl Acad. Sci. USA* **105,** 16411–16412 (2008).
51. Dray, N. *et al.* Hedgehog signaling regulates segment formation in the annelid Platynereis. *Science* **329,** 339–342 (2010).
52. Beaster-Jones, L. *et al.* Expression of somite segmentation genes in amphioxus: a clock without a wavefront? *Dev. Genes Evol.* **218,** 599–611 (2008).
53. Wagner, G. P. The developmental genetics of homology. *Nature Rev. Genet.* **8,** 473–479 (2007).
54. Northcutt, R. G. Evolution of centralized nervous systems: two schools of evolutionary thought. *Proc. Natl Acad. Sci. USA* **109** (Suppl. 1), 10626–10633 (2012).
55. van Wijhe, J. W. Über den vorderen neuroporus und die phylogenetische function des canalis neurentericus der wirbelthiere [in German]. *Zool. Anz.* **7,** 683–687 (1884).
56. Ziegler, H. E. Die phylogenetische entstehung des kopfes der wirbeltiere [in German]. *Jena. Zeitschr. Naturwiss.* **43,** 653–684 (1908).
57. Bjerring, H. C. Major anatomical steps toward craniotedness: a heterodox view based largely on embryological data. *J. Vert. Paleontol.* **4,** 17–29 (1984).
58. Hatschek, B. Studien über Entwicklungsgeschichte der Anneliden. Ein beitrag zur morphologie der Bilaterien [in German]. *Arb. Zool. Inst. Wien* **11,** 1–128 (1878).
59. Roule, L. Étude sur les forms premièrs de la notochorde et sur les affinitiés naturelles des cordés [in French]. *Arch. Zool. Exp. Gén.* (Sér.4) **10,** 447–547 (1909)
60. Delsman, H. C. *The Ancestry of Vertebrates* (Valkoff, 1922).
Feelings ran high about vertebrate-origin scenarios, for instance, J. W. van Wijhe thought this work "ought to be confiscated and consigned to the flames".
61. Marlow, H. *et al.* Larval body patterning and apical organs are conserved in animal evolution. *BMC Biol.* **12,** 7 (2014).
62. Garstang, W. Preliminary note on a new theory of the phylogeny of the Chordata. *Zool. Anz.* **17,** 122–125 (1894).
63. Jollie, M. The origin of chordates. *Acta Zool. Stockh.* **54,** 81–100 (1973).
64. Ivanova-Kazas, O. M. On the ancestry of Chordata and Deuterostomia as a whole. *Russ. J. Mar. Biol.* **23,** 219–226 (1997).
65. Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* (Oxford Univ. Press, 2001).
66. Kupffer, C. v. Die stammverwandtschaft zwischen Ascidien und wirbelthieren [in German]. *Ark. Mik. Anat.* **6,** 115–172 (1870).
67. Garstang, W. The morphology of the Tunicata, and its bearings on the phylogeny of the Chordata. *Q. J. Microsc. Sci.* **72,** 51–187 (1928).
68. Berrill, N. J. *The Origin of Vertebrates* (Clarendon, 1955).
69. Whitear, M. Some remarks on the ascidian affinities of vertebrates. *Ann. Mag. Nat. Hist.* (Ser. 12) **10,** 338–348 (1957).
70. Romer, A. S. *The Vertebrate Body* 3rd Edn (Saunders, 1962).
The scenario of Romer became well known because his book was read attentively by generations of American premedical students cramming for medical school acceptance.
71. Berg, L. S. *Nomogenesis or Evolution Determined by Law* (Constable, 1926).
72. Clark, A. H. *The New Evolution: Zoogenesis* (Williams & Wilkins, 1930).
73. Nursall, J. R. On the origin of the major groups of animals. *Evolution* **16,** 118–123 (1962).
74. Anderson, D. T. Origins and relationships among animal phyla. *Proc. Linn. Soc. N. S. W.* **106,** 151–166 (1982).
75. Sedgwick, A. On the origin of metameric segmentation and some other morphological questions. *Q. J. Microsc. Sci.* **24,** 43–82 (1884).
76. Lameere, A. L'origine des vertébrés [in French]. *Bull. Séanc. Soc. Belge Micros.* **17,** 91–121 (1891).
77. Inglis, W. G. Evolutionary waves: patterns in the origins of animal phyla. *Aust. J. Zool.* **33,** 153–178 (1985).
78. Dewel, R. A. Colonial origin for Eumetazoa: major morphological transitions and the origin of bilaterian complexity. *J. Morphol.* **243,** 35–74 (2000).
79. Tretjakoff, D. Ursprung der Chordaten [in German]. *Z. Wiss. Zool.* **134,** 558–640 (1929).
80. Christofersen, M. L. & Araujo-de-Almeida, E. A phylogenetic framework of the Enterocoela (Metameria: Coelomata). *Rev. Nordest. Biol.* **9,** 173–208 (1994).
81. Gerhart, J. & Kirschner, M. *Cells, Embryos and Evolution: Toward a Cellular and Developmental Understanding of Phenotypic Variation and Evolutionary Adaptability* (Blackwell, 1997).
82. Leydig, F. *Vom Bau des thierischen Körpers. Handbuch der vergleichenden Anatomie*, Vol. 1 [in German] (Laupp & Siebeck, Tübingen, 1864).
This is the first scenario for an invertebrate-to-vertebrate transition to appear after *On the Origin of Species* and to be written from a Darwinian point of view.
83. Gaudry, A. *Les Enchaînments du Monde Animal dans les Temps Géologiques* [in French] (Savy, 1883).
84. Jaeckel, O. Über die Stammform der Wirbelthiere [in German]. *Sitzungsber. Ges. Naturforsch. Freunde Berlin* **1896,** 107–129 (1896).
85. Gaskell, W. H. *The Origin of Vertebrates* (Longmans Green, 1908).

86. Raw, F. Outline of a theory of origin of the Vertebrata. *J. Paleontol.* **34,** 497–539 (1960).
87. Sillman, L. R. The origin of the vertebrates. *J. Paleontol.* **34,** 540–544 (1960).
88. Hoffhaus, C. E. A homogeneous theory of the origin of vertebrates. *J. Paleontol.* **37,** 458–471 (1963).
89. Løvtrup, S. *The Phylogeny of Vertebrata* (Wiley, 1977).
90. Bergström, J. The origin of animal phyla and the new phylum Procoelomata. *Lethaia* **22,** 259–269 (1989).
91. Hesse, R. *Tierbau und Tierleben in Ihrem Zusammenhang Betrachtet. Vol. I. Der Tierkörper als Selbständiger Organismus* [in German] (Teubner, 1910).
92. Naef, A. Notizen zur morphologie und stammgeschichte der wirbeltiere. 7. Das verhältnis der chordaten zu niederen tierformen und der typische verlauf ihrer frühen entwicklung [in German]. *Biol. Zentralbl.* **46,** 39–50 (1926).
93. Sepp, E. K. *Developmental History of the Nervous System of Vertebrates* [in Russian] (Medgiz, 1959).
94. Engelbrecht, D. V. Z. The annelid ancestry of the chordates and the origin of the chordate central nervous system and the notochord. *J. Zool. Syst. Evol. Res.* **7,** 18–30 (1969).
95. Gutmann, W. F. Relationships between invertebrate phyla based on functional-mechanical analysis of the hydrostatic skeleton. *Am. Zool.* **21,** 63–81 (1981).
96. Hubrecht, A. A. W. The relation of the Nemertea to the Vertebrata. *Q. J. Microsc. Sci.* **27,** 605–644 (1887).
97. Macfarlane, J. M. *The Causes and Course of Organic Evolution. A Study in Bioenergetics* (Macmillan, 1918).
98. Jensen, D. D. Hoplonemertines, myxinoids and deuterostome origins. *Nature* **188,** 649–650 (1960).
99. Willmer, E. N. Nemertines as possible ancestors of the vertebrates. *Biol. Rev. Camb. Philos. Soc.* **49,** 321–363 (1974).
100. Dzik, J. The origin of the mineral skeleton in chordates. *Evol. Biol.* **31,** 105–154 (2000).
101. Goette, A. Über den Ursprung der Wirbelthiere [in German]. *Verh. Dtsch. Zool. Ges.* **5,** 12–30 (1895).
102. Plate, L. Über den Ursprung der Wirbelthiere; eine kritische Besprechung [in German]. *Anat. Anz.* **58,** 39–46 (1924).
103. Salvini-Plawen, L. The urochordate larva and archicoelomate organization: chordate origin and anagenesis revisited. *J. Zool. Syst. Evol. Res.* **36,** 129–145 (1998).
104. Gregory, W. K. The transformation of organic design: a review of the origin and deployment of the earlier vertebrates. *Biol. Rev.* **11,** 311–344 (1936).
105. Dillon, L. S. The hydrocoel and the ancestry of the chordates. *Evolution* **19,** 436–446 (1965).
106. Eaton, T. H. The stem-tail problem and the ancestry of the chordates. *J. Paleontol.* **44,** 969–979 (1970).
107. Kuznetsov, A. N. Five longitudes in chordate body. *Theor. Biol. Forum* **105,** 21–35 (2012).
108. Béraneck, M. E. *Théories Récentes Sur la Descendance des Vertébrés* [in French] (Attinger, 1892).
109. Theophiloff, S. *Zur Phylogenie der Tunicaten: Eine kritische Studie* [in German]. PhD thesis, Univ. Jena (1892).
110. MacBride, E. W. A review of Professor Spengel's monograph on Balanoglossus. *Q. J. Microsc. Sci.* **36,** 385–420 (1894).
111. Kemna, A. L'origine de la corde dorsale [in French]. *Ann. Soc. Roy. Zool. Malacol. Belg.* **39,** lxxxv–clvii (1904).
112. van der Horst, C. J. Hemichordata in *Dr H. G. Bronn's Klassen und Ordnungen des Tierreichs* [in German]. Vol. 4, Part 4, Book 2, Section 2, Installments 1–5 (Akademische Verlagsgesellschaft, 1939).
113. Tokioka, T. Phylogenetic speculation of the Tunicata. *Publ. Seto Mar. Biol. Lab.* **19,** 43–63 (1971).
114. Cameron, C. B., Garey, J. R. & Swalla, B. J. Evolution of the chordate body plan: new insights from phylogenetic analysis of deuterostome phyla. *Proc. Natl Acad. Sci. USA* **97,** 4469–4474 (2000).
115. Röttinger, E. & Lowe, C. J. Evolutionary crossroads in developmental biology: hemichordates. *Development* **139,** 2463–2475 (2012).
116. Ayers, H. Concerning vertebrate cephalogenesis. *J. Morphol.* **4,** 221–245 (1890).
117. Willey, A. *Amphioxus and the Ancestry of the Vertebrates* (Macmillan, 1894).
118. Perrier, E. L'origine des vertébrés. *C. R. Acad. Sci. Paris* **126,** 1479–1486 (1898).
119. Manzanares, M. & Nieto, M. A. A celebration of the new head and an evaluation of the new mouth. *Neuron* **37,** 895–898 (2003).
120. Todaro, F. Sur l'origine phylogénétique des yeux des vertébrés et sur la signification des épiphises et des hypophyses de leur cerveau [in French]. *Arch. Ital. Biol.* **9,** 55–57 (1888).
121. Brooks, W. K. The genus Salpa. *Mem. Biol. Lab. Johns Hopkins Univ.* **2,** 1–303 (1893).
122. Delage, Y. & Hérouard, E. Traité de Zoologie Concrète. *Vol VII. Les Procordés* [in French] (Schleicher, 1898).
123. Sewertzoff, A. N. Directions of evolution. *Acta Zool.* **10,** 59–141 (1929).
124. Wada, H. Evolutionary history of free-swimming and sessile lifestyles in urochordates as deduced from 18S rDNA molecular phylogeny. *Mol. Biol. Evol.* **15,** 1189–1194 (1998).
125. Halanych, K. M. *et al.* Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science* **267,** 1641–1643 (1995).
This pioneering study in molecular phylogenetics indicated that the greater part of the animal kingdom is divisible into three major super-phyletic groups.

The deuterostome context of chordate origins

Christopher J. Lowe[1], D. Nathaniel Clarke[1], Daniel M. Medeiros[2], Daniel S. Rokhsar[3,4,5] & John Gerhart[3]

Our understanding of vertebrate origins is powerfully informed by comparative morphology, embryology and genomics of chordates, hemichordates and echinoderms, which together make up the deuterostome clade. Striking body–plan differences among these phyla have historically hindered the identification of ancestral morphological features, but recent progress in molecular genetics and embryology has revealed deep similarities in body–axis formation and organization across deuterostomes, at stages before morphological differences develop. These developmental genetic features, along with robust support of pharyngeal gill slits as a shared deuterostome character, provide the foundation for the emergence of chordates.

The mystery of chordate origins has endured for more than 150 years. Shortly after Darwin's *On the Origin of Species*, acorn worms were discovered to have chordate-like pharyngeal gill slits[1,2] and to metamorphose from echinoderm-like larva[3], thus linking the evolution of chordates, hemichordates and echinoderms. Modern phylogenetic analysis has confirmed the union of these three phyla in a single clade. This group, the deuterostomes, provides the phylogenetic framework for developing hypotheses about the origin of chordate features through comparative morphology, embryology and genomics.

The emergence of comparative molecular developmental biology over the past quarter of a century has revived interest in classic hypotheses of animal body-plan evolution[4]. The comparative approach focuses on identifying morphological, developmental and genetic traits that are shared across phyla by virtue of their inheritance from a common ancestor, and provides an understanding of how such ancestral traits can arise and be subsequently modified. Although many recent hypotheses on chordate and vertebrate origins on the basis of molecular data are motivated primarily by projections from the bilaterian ancestor[4,5], a growing body of data from hemichordates, echinoderms and invertebrate chordates serves as the foundation for new hypotheses based on deuterostome ancestral characters[6–14].

Despite the impressive morphological disparity among deuterostome phyla, we are making progress identifying conserved anatomical and molecular ancestral characters. Each phylum is a fascinating natural experiment in body-plan evolution, but their dazzling diversity presents a major challenge for reconstructing early deuterostome evolutionary history in morphological terms (Box 1)[15]. In this Review we highlight recent advances in deuterostome phylogenetics, developmental biology and genomics that have contributed to our understanding of the early evolution of deuterostomes and the subsequent origin of chordates.

Deuterostome phylogeny

The first step in unravelling chordate origins is the establishment of a robust deuterostome phylogeny (Fig. 1). The chordates, uniting vertebrates, tunicates and cephalochordates, were first recognized by Haeckel[16], partly based on shared developmental characteristics. A key insight came from Kowalevsky's[17] recognition that the tadpole larva of ascidians shared many characteristics with vertebrates, an observation that greatly

impressed Darwin[18]. Kowalevsky also recognized the vertebrate-like gill slits of the invertebrate acorn worms[2]. The link between chordates and acorn worms was emphasized by Bateson, who proposed further morphological affinities between them in the late 1800s, and named the acorn worms 'hemichordates'[1]. Around the same time, Metchnikoff recognized the similar larval forms of hemichordates and echinoderms, and united these two phyla into the 'Ambulacraria'[3] (Box 2).

The unity of chordates, hemichordates and echinoderms was inferred by Grobben[19] on the basis of three shared developmental features: 'deuterostomous' development (derivation of the mouth from a secondary opening rather than the blastopore), radial cleavage and enterocoely (the pouching out of mesoderm from the archenteron wall). Although he named this lineage the 'deuterostomes' (second mouth), we now recognize that these features are not unique to the chordate–hemichordate–echinoderm clade, and are found in several other phyla[20], the result of either shared ancestry or convergence. This leads to the nomenclatural embarrassment that some phyla with deuterostomous development are not deuterostomes. Nevertheless the name has stuck, and by convention we refer to the chordate–hemichordate–echinoderm clade as the deuterostomes.

The advent of molecular phylogenomics has brought new methods to bear on the relationships between and within deuterostome phyla (Fig. 1). Ambulacraria, the surprising grouping of hemichordates and echinoderms, is strongly supported by molecular characters[15,21–23], and is clearly the sister group of chordates. Within chordates, it is now widely recognized that the cephalochordate lineage (amphioxus) diverged before the split between tunicates and vertebrates[21,24]. This recent discovery overturned earlier thinking that tunicates diverged first, which had implied that the simple ascidian tadpole larva represents ancestral chordate features (Box 1).
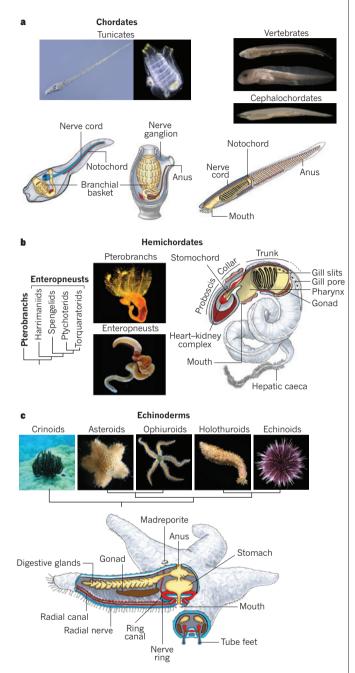
Although classic embryological criteria suggested that lophophorates (phoronids, brachiopods and bryozoans) and/or chaetognaths should also be grouped among the deuterostomes, molecular phylogenetics robustly supports their position in the protostomes[22,25,26]. Xenoturbellid worms are a more challenging case: these animals resemble acoelomorphs (acoel flatworms and nematodermatids) and have been grouped with them in a 'Xenacoelomorpha' clade[27,28]. Some molecular analyses also identify *Xenoturbella* and its relatives as ambulacrarians, and therefore deuterostomes[27], whereas other studies find that acoelomorphs diverge from

[1]Hopkins Marine Station, Department of Biology, Stanford University, 120 Oceanview Boulevard, Pacific Grove, California 93950, USA. [2]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder Ramaley N122 Campus Box 334 Boulder, Colorado 80309-0334, USA. [3]Department of Molecular and Cell Biology University of California, Berkeley 142 Life Sciences Addition 3200 Berkeley, California 94720-3200, USA. [4]Okinawa Institute of Science and Technology 1919-1 Tancha, Onna, Okinawa Prefecture 904-0495, Japan. [5]Department of Energy Joint Genome Institute 2800 Mitchell Drive Walnut Creek, California 94598, USA.

BOX 1

Deuterostome diversity

Unambiguous homologies between deuterostome phyla with morphologically disparate body plans are difficult to establish, leading to a wide range of often contradictory hypotheses about chordate origins[1,6,29,96,100,107,108,114–116]. We present basic descriptions of the adult body plans of the uncontested deuterostome phyla: chordates, hemichordates and echinoderms. All mesodermally derived-structures are red, ectoderm are blue and endoderm are yellow. **a**, Chordates are set apart from other deuterostomes by a suite of features that enable swimming by paired muscles along a trunk that extends post-anally. These muscles exert forces on the notochord, a flexible rod that provides elastic recoil to power movement. Chordates also have a unique tubular central nervous system (CNS)[29]. Of the subphyla, vertebrates are distinguished from other chordates by the elaboration of the head region with an enlarged anterior CNS with paired sense organs, evident here in a lamprey ammocoete larva (top) and an axolotl tadpole (bottom). Tunicates (larvaceans, ascidians and thalacians) are a diverse group of marine filter feeders that display a range of body plans and life-history strategies, including solitary, colonial, sessile and free-swimming forms[117]. They are represented here by ascidians. Chordate affinities are most evident in the larval form: an ascidian tadpole (left) has a tubular nerve cord, a notochord and a post-anal tail. These features regress at metamorphosis, leaving the branchial basket, a small nerve ganglion and the endostyle as the only chordate characters remaining in the adult (right). Cephalochordates, represented by amphioxus, are filter feeders that burrow in sand with their mouths open to the water column. Amphioxus shares much of its basic anatomy with vertebrates, including segmented musculature, and a vertebrate-like heart and circulatory system[30,118]. They have a modest CNS consisting of a neural tube with simplified vertebrate-type patterning along both the anteroposterior and dorsoventral axes[8,118,119]. **b**, Hemichordates are a clade of marine worms divided into two groups: enteropneusts and pterobranchs. Hemichordate phylogeny is based on Cannon *et al.*[120]. Pterobranchs, shown here by *Cephalodiscus*, are small largely colonial animals that live within the protection of a secreted fibrous tube and use a ciliated lophophore for filter feeding[50,58]. Enteropneusts, or acorn worms, are solitary, burrowing worms that feed using a combination of deposit and filter feeding[52,121]. The harrimaniid *Saccoglossus kowalevskii*, which has been used for many developmental studies[12], is pictured (micrograph). Both groups of hemichordates are united by their tripartite body plan, which includes proboscis, collar and trunk (as shown in the illustration of a spengelid entropneust). The proboscis is used for digging and feeding and contains the gut diverticulum called the stomochord that supports a heart–kidney complex[56,60]. The mouth opens ventrally into the pharynx within the collar region, and the anterior trunk is perforated by a series of dorsolateral gill slits[58]. **c**, Echinoderms have considerably modified the ancestral bilaterian body plan to become pentaradially symmetrical as adults, although their larvae are bilaterally symmetric (Box 2). Even basic axis comparisons with other deuterostomes are problematic, and the evolutionary origins of this phylum remain a mystery. All five extant classes of echinoderms: crinoids (sea lilies), asteroids (sea stars), ophiuroids (brittle stars), holothuroids (sea cucumbers) and echinoids (sea urchins) are characterized by a conserved body plan shown by a diagram of an asteroid with cutaways to show internal anatomy; the

a Chordates
Tunicates
Vertebrates
Cephalochordates

Nerve cord · Notochord · Branchial basket · Nerve ganglion · Anus · Notochord · Nerve cord · Anus · Mouth

b Hemichordates
Pterobranchs
Enteropneusts
Pterobranchs · Harrimaniids · Spengelids · Ptychoterids · Torquaratorids
Stomochord · Collar · Trunk · Proboscis · Gill slits · Gill pore · Pharynx · Gonad · Heart–kidney complex · Mouth · Hepatic caeca

c Echinoderms
Crinoids · Asteroids · Ophiuroids · Holothuroids · Echinoids
Madreporite · Anus · Digestive glands · Gonad · Stomach · Radial canal · Radial nerve · Ring canal · Nerve ring · Mouth · Tube feet

mesodermally derived water vascular system, a hydraulic system that drives the distinctive tube feet used for feeding and locomotion; five radial nerves along each arm/ambulacrum linked by a nerve ring, and the mesodermally derived skeleton. Asteroids most clearly exhibit the basic components of the body plan. Phylogenetic relationships are based on refs 120, 122.

the bilaterian stem before the protostome–deuterostome split[28] (Fig. 1). We note, however, that even if xenoturbellids and/or acoelomorphs are deuterostomes, their simple body plans would represent secondary loss from a more complex deuterostome ancestor. The resolution of the phylogenetic placement of these taxa is therefore unlikely to provide substantial insight into vertebrate origins.

Ancestral chordate characters

On the basis of shared features of living chordates we have gained a rather detailed view of the development, morphology and life history of the last common chordate ancestor. Most classic and modern reconstructions of ancestral chordates propose a filter feeder with a notochord, gill slits, endostyle, dorsal hollow nerve cord and post-anal tail[29]. The recent

revision of the chordate family tree has added to this list of ancestral chordate features. The basal position of cephalochordates among chordates suggests that similarities between amphioxus and vertebrates represent ancestral chordate features lost in tunicates[21,24]. Thus, in addition to the core features listed earlier, the Early Cambrian or Pre-Cambrian chordate ancestor probably possessed myomeres, a vertebrate-like circulatory system and a central nervous system (CNS)[30]. The life history of cephalochordates, and the fact that larval lampreys and adult hemichordates are burrowing filter feeders, further suggest that this ancestor was a solitary, endobenthic filter feeder that was capable of short swims.

The striking similarities between amphioxus- and vertebrate-developmental mechanisms allow a fairly comprehensive reconstruction of early development in primitive chordates. As in vertebrates and cephalochordates, the anteroposterior (AP) and dorsoventral (DV) axes of the ancestral chordate were probably determined during gastrula stages by organizing centres much like Spemann's organizer of vertebrates, secreting long-range patterning signals[11]. Opposing Nodal and BMP signalling gradients established the DV axis, with Chordin-mediated BMP inhibition in the dorsal ectoderm segregating the presumptive CNS from the epidermal (or general) ectoderm[11,31]. Along the AP axis, Wnt and retinoic acid signalling probably acted on Hox genes and other transcription-factor genes to establish the regional identities of AP domains of the body axis, including the boundary between the foregut and hindgut and the main subdivisions of the CNS[8,32,33].

Comparisons between amphioxus and vertebrates suggest a deep ancestry of the major divisions of the CNS along the AP axis. Later in development, fine-scale patterning of the ancestral chordate CNS was also vertebrate-like, but simpler. Along the DV axis of the CNS, all chordates have a molecularly distinct dorsal domain that expresses *pax3/7*, *msx* and *zic* genes and generates sensory interneuron cells[34], a ventral floor plate expressing hedgehog ligands[35], and an intervening bilateral domain flanking the neural tube lumen and generating motor and visceral neurons. The expression domains of transcription factors and signalling molecules along the AP axis of the CNS are also mostly conserved across chordates, and presumably reflect expression domains of the chordate ancestor[8]. Precisely how this patterning was generated is less clear, as current data suggest that neither amphioxus nor tunicates have unambiguous, functionally validated homologues of two vertebrate CNS signalling centres, the isthmic organizer or the zona limitans intrathalamica (although these signalling mechanisms may have been present in a deuterostome ancestor, see later)[8,36,37].

Segmented musculature of the ancestral chordate almost certainly developed from somites, and at least some formed by enterocoely[35,38]. In amphioxus, the anterior-most somites form by enterocoely, whereas posterior somites pinch off sequentially from the tail bud[36,39]. In vertebrates, a 'clock and wavefront' mechanism, involving oscillating Notch and Wnt

signalling and a posterior fibroblast growth factor (FGF)-signalling gradient divides the paraxial mesoderm into a series of somites[40]. Despite these mechanistic differences, amphioxus displays vertebrate-like segmental expression of Notch and Wnt signalling components in nascent somites, and requires FGF signalling for forming and maturing the anterior and posterior somites[41,42]. Thus, somitogenesis in all living chordates, and presumably their last common ancestor, involved iterated Notch–Delta and Wnt signalling, and FGFs.
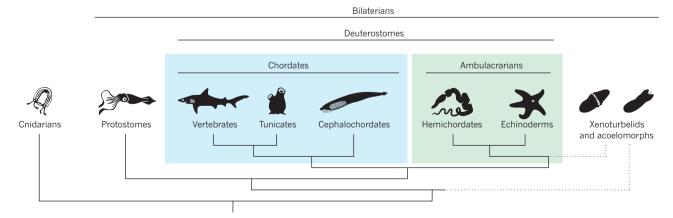
Despite differences in when and how the pharyngeal gill slits form in the three chordate clades, recent work reveals conserved aspects of their development, presumably inherited from the chordate common ancestor. In amphioxus and vertebrates, the pharyngeal endoderm is specified by attenuated retinoic acid signalling, and marked by conserved expression of several transcription factors including pax1/9, six1/2, six4/5, six3/6, eya, foxC and foxL1 (refs 32, 43, 44). In addition, recent work has shown that the chordate ancestor probably had a collagen-based pharyngeal skeleton incorporating cellular[45] and acellular cartilage[46,47] derived from pharyngeal mesoderm. Whereas the pharyngeal walls develop pharyngeal pouches and gill slits, the floor develops endostyle specializations related to trapping food particles during filter feeding, as well as to hormonal and protective functions.

The deuterostome roots of chordate characteristics

Work on hemichordates and echinoderms has informed our understanding of ancestral deuterostome features, with different taxa contributing complementary insights. Integrating insights from echinoderms is challenging owing to the divergent radial body plan of adults, although studies of echinoderm larval development have made essential contributions to our understanding of early deuterostome embryogenesis[13,14,48,49]. Pterobranch hemichordates are relatively understudied[50]. In this Review, therefore, we focus primarily on insights derived from the study of enteropneust hemichordates (acorn worms) as they relate to our understanding of early deuterostome evolution.

As first described by Kowalevsky[2], the anterior gut of hemichordates is perforated in the dorsolateral region by a series of ciliated gill slits, now known to be supported by gill bars composed of an acellular collagen secreted by the endoderm (Fig. 2a, Box 1)[46]. Although there is no equivalent structure in extant echinoderms, fossils reveal compelling evidence that gill slits were present in stem echinoderms and subsequently lost[51]. On the basis of morphological and functional criteria, enteropneust gill slits closely resemble those of cephalochordates and are plausibly homologous[1,46,52]. In two species of enteropneust, studies of patterning genes with conserved roles in chordate gill-pouch development, namely *pax1/9*, *foxC*, *foxL1*, *eya*, *six1* and *foxI*, also strongly support homology[53–55].

The stomochord in hemichordates has drawn much comparative interest as a notochord-like ancestral trait[1,29,56,57]. It is a diverticulum of

Figure 1 | Deuterostome phylogeny. A consensus cladogram of deuterostome groups based on recent phylogenomic data sets[21,22,24,28,113]. There are three major phyla of extant deuterostomes, which are grouped into two diverse clades: the ambulacrarian phyla (green), consisting of hemichordates and echinoderms, and chordates (blue), consisting of the cephalochordate, tunicate and vertebrate lineages. Recent analyses have proposed either a grouping of xenoturbellid and acoelomorph flatworms as sister group to ambulacrarians[27], or at the base of the bilaterians[28] (dashed lines).

Figure 2 | Key anatomical features of the enteropneust body plan.
a, Longitudinal and transverse sections through an adult enteropneust hemichordate, highlighting morphological characters that have featured prominently in classic hypotheses of deuterostome evolution and chordate origins. A, anterior; P, posterior; D, dorsal; V, ventral. **b**, The nervous system of an adult enteropneust showing both the broad basiepithelial plexus throughout the ectoderm and nerve chords along the dorsal and ventral midlines. The blue spots represent cell bodies and the lines represent neural processes.

the anterior gut that extends into the posterior proboscis supporting the heart–kidney complex on its dorsal surface (Fig. 2a, Box 1). Stomochord cells are vacuolated and surrounded by a sheath, similar in tissue organization to a notochord[1,56,58]. However, homology of these two structures is weakly supported by both morphological and molecular criteria[59–61]. In chordates the developing notochord is a key source of the secreted BMP antagonists Chordin, Noggin and Follistatin, and the ventralizing ligand Shh[62]. Of these genetic markers, only *hh* (the homologue of *Shh*) is expressed in the stomochord, but it is also observed in surrounding anterior endoderm[57,63]. Possible alternatives to notochord homology are suggested by the stomochord expression of genes such as *otx*, *dmbx*, *hex* and *foxE* that are expressed in prechordal endomesoderm of chordates, but not in the notochord. These markers suggest that the stomochord is an anterior endodermal structure with stronger affinities to the endostyle than the notochord[61].

The hemichordate nervous system is characterized by two contrasting organizational features (Fig. 2b): a broad basiepithelial plexus, particularly prominent in proboscis ectoderm, and a pair of nerve cords. The ventral cord extends the length of the trunk and the dorsal cord runs from the base of the proboscis down the length of the animal and joins to the ventral cord by lateral nerve rings. Both cords are superficial condensations of the nerve plexus except in a short length that spans the collar, where the cord is internalized into a tube with a prominent lumen in some species, and is formed by a developmental process that resembles chordate neurulation[59,64–67]. Various authors have proposed both cords as possible homologues of the chordate dorsal cord[57,59,68,69], however, the internalized collar cord has attracted the most attention[6,57,59,69]. Early reports suggested that the dorsal cord was simply a through conduction tract of axons[70,71]. Molecular studies, however, have shown condensations of cell bodies associated with this cord[6,69], and a further study in *Balanoglossus simodensis* revealed *bmp2/4*, *pax3/7* and *msx* expression in the collar cord[57], similar to that of the most lateral parts of the vertebrate neural plate and in other bilaterians during CNS development. Although these similarities are supportive of homology of the collar cord and chordate nerve cord, other neural molecular markers complicate this interpretation. In *Saccoglossus kowalevskii*, markers of medial rather than lateral neural plate are not expressed in the dorsal cord as predicted, but rather along the ventral midline associated with the ventral cord. In addition, several neural markers are not only expressed in the collar cord, but also throughout the length of the superficial cord in the trunk, suggesting a patterning role

throughout the dorsal midline[72]. When considering the general organization of the nervous system in enteropneusts, no simple homology statements can yet be made in relation to other nervous systems.

Although it seems likely that ancestral deuterostomes inherited some elements of nervous system centralization from the bilaterian common ancestor, a comprehensive characterization of key molecular markers is needed to test competing hypotheses of nervous system evolution further. It remains unclear whether the main features of the unusual enteropneust nervous system can be ascribed to the filter-feeding deuterostome ancestor, thereafter modified in the chordate line, or whether they are secondary derivatives of the hemichordate lineage.

Axial patterning of deuterostome body plans
The discovery of conserved, pan-bilaterian mechanisms for the development of the animal–vegetal, AP and DV body axes has transformed our thinking about animal evolution[4,5,73]. This deep conservation initially surprised biologists because of the great morphological diversity of bilaterians, but made more sense when it was realized that the early axiation processes of the embryo are separate from the later processes of morphogenesis, organogenesis and cell differentiation. Conserved suites of genes are responsible for establishing basic regional differences of cells along all three axes of bilaterian embryos, reflecting an extensive genetic regulatory network spread across the developing embryo. The resulting map of conserved expression domains represents an 'invisible anatomy'[74] that reveals clear relationships between disparate body plans, and provides a window into the organization of expression domains in the deuterostome ancestor. In this Review, we focus on the mechanisms by which these axes are formed in deuterostomes, and the patterns of transcription-factor and signalling-gene-expression domains produced along these axes.

The animal–vegetal axis and formation of endomesoderm
One of the first developmental decisions in embryogenesis is the establishment of the animal–vegetal axis. This axis sets up the formation of the three germ layers: endoderm, mesoderm and ectoderm. Ectoderm derives from the animal pole, and endomesoderm from the vegetal pole, which later divides into endoderm and mesoderm. In all three major deuterostome phyla, the formation of endomesoderm is triggered by β-catenin protein, the intracellular effector of the canonical Wnt signalling pathway. β-Catenin is stabilized preferentially in the vegetal pole of early embryos and activates genes of the endomesodermal cellular program[75,76].
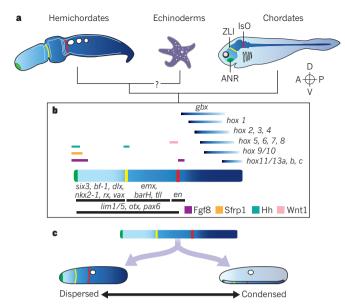
Figure 3 | A conserved molecular network for the deuterostome anteroposterior axis. a, Schematic representation of the distribution of ectodermal expression domains of anteroposterior (AP) transcription factors (blue gradient) and ectodermal signalling centres (green, yellow and red) in relation to the body plans of deuterostome phyla. Chordate neuroectodermal signalling centres depicted are the anterior neural ridge (ANR), zona limitans intrathalamica (ZLI) and isthmic organizer (IsO). Broad conservation of expression domains between hemichordates and chordates allows for the reconstruction of an ancestral patterning network, which is shown without any explicit inference of ancestral morphologies (**b**). Insufficient data exist from echinoderms to infer to what extent they share this conserved AP patterning network during adult patterning, although much of the anterior network is conserved in larvae[13,49]. **b**, Domain map for the conserved transcription factors and signalling ligands in relation to the AP axis[63,85,86]. **c**, Current data allow for the reconstruction of a conserved molecular coordinate system for the AP axis of the last common deuterostome ancestor, but not for the reconstruction of discrete morphologies of that ancestor, because this AP patterning network is deployed in a variety of morphological contexts, as evidenced by comparative data from hemichordates (dispersed; AP expression domains encircling the body) and chordates (condensed; AP domains largely restricted to regions near the dorsal midline). A, anterior; P, posterior; D, dorsal; V, ventral.

In hemichordates and echinoderms, knockdown of the gene that encodes β-catenin protein results in the 'animalization' of the embryo — excess ectoderm and no endomesoderm. Conversely, stabilization of the protein throughout the embryo results in 'vegetalization' of the embryo — excess endomesoderm and no ectoderm. This mechanism has also been demonstrated in protostomes from work on nemertine embryos[77], and β-catenin protein is also involved in endoderm formation in cnidarians[78], suggesting a deep eumetazoan ancestry for this process[73].

Later specification of mesoderm from the endomesoderm occurs by either of two generic mechanisms: autonomous specification by a cell's inheritance of a sequestered cytoplasmic determinant, or induction by a signal from neighbouring tissue. In all deuterostomes except ascidians, mesoderm formation occurs by induction. In vertebrates, two main signalling pathways are involved in mesoderm specification: Nodal and FGF[79]. In amphioxus, FGF signalling specifies anterior mesoderm that forms by enterocoely[42]. Similarly, in the hemichordate *S. kowalevskii*, FGF signalling induces mesoderm and enterocoely, which raises the possibility of an ancestral role of FGF in deuterostome mesoderm formation[80]. As a classic deuterostome character, a mechanistic link of enterocoely to FGF signalling would support homology of this trait, at least within the deuterostomes. In echinoderms, however, the role of FGF has yet to be fully characterized, and there is some variation in inductive cues involved in mesoderm specification: Notch–Delta signalling is important in early mesoderm specification of echinoids, but not asteroids[81]. The differences

between deuterostomes in specifying endomesoderm and mesoderm preclude the definitive inference of the pathway of the deuterostome ancestor, except that β-catenin protein is required at the start, and various inductive signals are required later.

Anteroposterior axis

Although deuterostome taxa show an impressive array of morphologies, organs and cell types along the AP axis, many of the early developmental steps of axis formation are highly conserved and probably date back to the bilaterian ancestor. Wnt signalling through β-catenin has emerged as the earliest conserved determinant of AP pattern in deuterostomes. (Note that this time and place of usage of β-catenin is separate from its role in endomesoderm formation discussed earlier.) In vertebrates, Wnt proteins act as posteriorizing signals in all three germ layers, but are most analysed in CNS patterning[82,83]. Whereas Wnts are produced posteriorly, Wnt antagonists are produced anteriorly from the mesoderm of Spemann's organizer, and their interaction sets up a graded Wnt distribution prefiguring the eventual anatomical AP axis[84]. In both sea urchin larvae and the direct-developing *S. kowalevskii*, Wnt signalling is also important for establishing AP patterning[48,63,75], suggesting that generating a Wnt signalling gradient (high posteriorly, low anteriorly) is a key step in AP-axis formation in all three phyla, for both adult and larval body plans. Different intensities of Wnt signalling along the graded distribution then activate distinct genes encoding different transcription factors and signalling ligands, producing a long-lasting AP map of gene expression domains that is collinear with the Wnt distribution. The ectodermal map is strikingly similar in the identity and relative expression of the constituent regulatory genes across bilaterians[85,86].

This conserved AP map provides a novel basis for comparing body plans (Fig. 3a, b)[74]. In the most anterior regions, coexpression of genes such as *sfrp1/5*, *fgf8/17/18*, *foxG*, *retinal homeobox*, *dlx* and *nk2-1* define ectodermal territories that later form proboscis ectoderm in hemichordates and forebrain in vertebrates. Further posteriorly, expression domains of *emx*, *barH*, *dmbx* and *pax6* define the collar ectoderm of hemichordates and midbrain of vertebrates; still more posteriorly, domains of *gbx*, *engrailed*, *pax2/5/8* and the collinearly expressed Hox genes, regulate pharynx and trunk patterning of hemichordates and the hindbrain and spinal cord in vertebrates (Fig. 3a)[63,85,86]. Enteropneust Hox genes are organized as an intact cluster[87], and in both cases the posterior group Hox genes are expressed in post-anal parts of the body axis, perhaps indicating domain-level homology of these deuterostomian posterior appendages. AP map similarities even extend to three signalling centres, producing the same signals and occupying equivalent map positions, that are important for vertebrate brain patterning and for hemichordate ectodermal development at the anterior tip, proboscis–collar boundary and collar–trunk boundary (Fig. 3a,b) (for an alternative perspective see ref. 88). In hemichordates, the conserved AP map of ectodermal expression domains covers both neural and epidermal tissue, and domains encircle the body. In chordates most comparative studies have focused on the role of this network in patterning the dorsal CNS, but more recent studies demonstrate that expression of many of the genes extend ventrally into sensory neurons and epidermis, suggesting a more general role in ectodermal patterning[88].

The AP map of expression domains provides a positional criterion for evaluating morphological homologies between disparate body plans. Thus, the homology of chordate and hemichordate gill slits is supported by the observation that in both groups the first slit perforates the same region of the AP expression map, near the midbrain–hindbrain boundary in chordates and the collar–trunk boundary in enteropneusts. The map also provokes comparisons: if the hemichordate collar cord is homologous to the chordate dorsal nerve cord, it should express AP genes similar to those of the chordate midbrain. We can confidently reconstruct this AP patterning network in the ectoderm of the deuterostome ancestor, and as previously mentioned, much of the map probably dates back to the bilaterian ancestor. Indeed, more comparisons with protostomes are needed to illuminate which few domains are deuterostome-unique, for example,
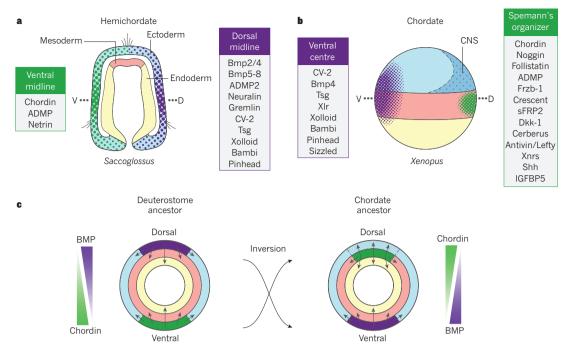
Figure 4 | Comparison of the dorsoventral patterning mechanisms of hemichordates and chordates. a, BMP–Chordin signalling components expressed in the dorsal and ventral midline ectoderm (blue) in the late gastrula stage of *Saccoglossus kowalevskii*. **b**, BMP–Chordin signalling components expressed either on the ventral side or dorsally in Spemann's organizer in the early gastrula of *Xenopus*. CNS, central nervous system. **c**, The inversion of dorsoventral (DV) signalling centres and the relocation of the Chordin source from the ectoderm (yellow) to mesoderm (red) were innovations in DV patterning at the base of the chordates (ancestral location shown by grey shading).

in pharyngeal ectoderm and endoderm.

The AP axial homology of chordates and hemichordates with echinoderms is far less clear[89–93]. During the development of the larvae of asteroids, echinoids and crinoids (Box 2), anterior regulatory genes are expressed throughout the anterior ectoderm[49,94], whereas posterior markers such as Hox genes are entirely absent during early patterning. In both echinoids and crinoids, it is not until adult patterning begins in the late larva that Hox genes begin expression in a collinear pattern, not in ectoderm, but in posterior coelomic mesoderm[90,95]. Some anterior markers are expressed in the oral ectoderm and tube feet of pentaradial adults, but current data are too fragmentary to make valid comparisons of adult echinoderms with other deuterostome adults. Comprehensive characterization of the patterning of echinoderm adults is badly needed to unravel the evolution of this unusual body plan.

Although the AP map is conserved across deuterostomes (and in most aspects, across bilaterians) the differentiated morphologies built on it are probably not (Fig. 3c). The morphological outcomes of development differ in each phylum because the transcription factors and signals of the conserved map activate and repress different target genes[63]. These target genes, in turn, direct the final steps of organogenesis, morphogenesis and cell-type formation.

The dorsoventral dimension

The DV axis evolved on the Pre-Cambrian stem leading to the bilaterian ancestor, and is intimately tied with the origin of bilateral symmetry. Its formation in early embryogenesis is analogous to AP axis formation. One midline of the embryo produces Bmp, and the opposite midline produces the Bmp antagonist Chordin[96]. Through complex interactions, this antagonism generates a graded distribution of Bmp across the embryo, a graded occupancy of Bmp receptors, and a corresponding graded distribution of activated Smad1/5 transcription factor in embryonic cells. This gradient of activated Smad1/5 stimulates and represses different genes encoding transcription factors and other signalling ligands, generating a long-lasting DV map of expression domains of these genes[5,97] (Fig. 4a, b).

The patterns of transcription-factor and signalling-ligand expression established along the DV direction generate the corresponding anatomical axis by driving the expression of genes for the development of different tissues and cell types in different regions. Some of the definitive tissues and cell types are remarkably conserved among bilaterians, as demonstrated by the similarities between the DV development of protostomes such as the fruit fly and the annelid *Platynereis dumerilii* and vertebrates such as *Xenopus*, mice and zebrafish[97,98]. Domains from the Chordin side of the Bmp distribution activate axial (striated) muscle development in the mesoderm and nerve-cell development in the ectoderm, especially motor neurons and interneurons that assemble into the CNS, whereas domains from the Bmp side activate heart tube and coelom development from the mesoderm and epidermis and sensory-nerve-cell development from the ectoderm[97]. The Bmp distribution patterns all three germ layers.

Although deuterostomes as a group inherited the basic mechanism of DV axis formation from the bilateral ancestor, there are important differences among them that can inform hypotheses of chordate origins. It is immediately apparent that the Chordin and Bmp sides of the molecular DV axis have different anatomical names in deuterostomes and protostomes. In deuterostomes, the Bmp side is called 'dorsal' and the Chordin side is 'ventral', but in *Drosophila* and other protostomes the molecular and anatomical links are reversed. By zoological convention, sides are named according to the animal's orientation to the substratum and the location of the mouth. The difference was resolved by the proposal that the chordate ancestor underwent a dorsoventral inversion of the body relative to the substratum. This transition simultaneously inverted the Bmp–Chordin axis, the domain map, and axis of anatomical differentiations[5,96]. As a final refinement the mouth was relocated to the Bmp side, whereas most protostomes (for example, *Drosophila*), and invertebrate deuterostomes, form the mouth on the Chordin side. Although seeming modest as a novelty, body inversion must be considered when discussing innovations of the chordate line.
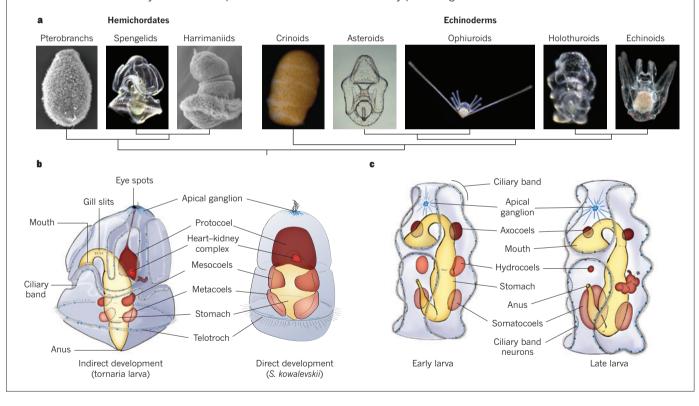
S. kowalevskii provides an excellent example of bilaterian DV axiation, probably conserved from the deuterostome ancestor[72]. (Indirect developing hemichordates and echinoderms also exhibit Bmp–Chordin-based DV patterning, modified for larval body plans, although we cannot cover these here[9,99].) At gastrula stages, *bmp2/4* is strongly expressed on the dorsal ectodermal midline of *S. kowalevskii*, accompanied by genes for a large

BOX 2

Deuterostome larval diversity

a, Hemichordates and echinoderms include lineages that are characterized by both direct-developers, forming the adult body plan from embryogenesis in a matter of days; and by indirect-developers, first forming planktonic feeding larvae that may swim and feed for months before metamorphosing to produce a distinct adult body plan. The harrimaniid enteropneusts and pterobranchs (reproduced from ref. 49) are examples of direct-developers. **b**, A model of a two-day-old embryo of *Saccoglossus kowalevskii* contrasts with the month-old late spengelid tornaria larva. **c**, The organization of the tornaria larval body plan is very similar to the organization of the echinoderm larva represented here by a holothuroid auricularian larva. However, echinoderms have a spectacular variety of larval forms from the ophiopluteus and echinopluteus with similar elaborate skeletons to the asteroid bipinnarian and holothuroid auricularian larvae with similar convoluted ciliary bands. Many researchers have focused on the morphological and developmental similarities between the diverse ambulacrarian larval types, suggesting the existence of an ancestral 'dipleurula' (small two-sided) larval form from which ambulacrarian larval diversity arose[114]. The dipleurula ectoderm is

characterized by a convoluted ciliary band used for swimming and feeding. In hemichordates, a robust additional posterior band of compound cilia, the telotroch, is purely locomotory (**a**, **b**). The nervous system is divided into two domains: an apical ganglion underlying the sensory ciliated apical organ, and neurons underlying the length of the ciliary bands (**b**, **c**). The dipleurula mesoderm is formed by enterocoely and organized into three compartments: anterior, middle and posterior. In echinoderms, the adult body plan is initiated by the left middle coelom, which expands and forms five lobes midway through larval development (**b**, asterisk in **c**). An influential theory of Garstang[108] further elaborated by a variety of authors (see Review by Holland *et al*. on page 450), proposed that the deuterostome ancestor also had a dipleurula larva, and that chordates evolved by paedomorphosis from such forms. A central tenet of this theory is that the dorsal central nervous system of chordates evolved through the dorsal migration and fusion of the lateral ciliary bands of the dipleurula larvae, and their underlying neurons[114]. More recently, this hypothesis has fallen out of favour on the basis of both phylogenetic and body-patterning data[21,24,109].

set of signal modulating proteins and other Bmp-related proteins. Conversely, *chordin* and *admp* are strongly expressed on the opposite, ventral midline (Fig. 4a). Following the Bmp distribution gradient, transcription-factor genes are activated in a DV map that generally parallels the expression of orthologous genes in *Drosophila* and vertebrates. The DV domain map and subsequent differentiated structures of the overt anatomical axis depend entirely on the Bmp distribution, as shown by the development of dorsalized embryos in the presence of excess uniform Bmp2/4 protein, and of ventralized embryos when Bmp2/4 is eliminated[72]. Tissues, organs and cell types of the three germ layers are patterned by the Bmp–Chordin distribution, including the gill slits, the mouth and the two nerve cords (Fig. 2a). In embryos dorsalized by excessive Bmp, nerve cells still form in abundance. Although this might seem contrary to chordate neural

patterning in which Bmp initially represses neural development in the epidermis, it is not; the hemichordate dorsal nerve cord normally forms at the midline of high Bmp concentration, and the lateral parts of the chordate neural plate are themselves patterned by high Bmp concentrations. Overall, the hemichordate findings affirm general insights about bilaterian DV axis formation. In its body orientation, *S. kowalevskii* resembles protostomes: Bmp foretells the ventral side and Chordin the dorsal, leaving chordates as the single 'inverted' phylum (Fig. 4c). One of the key questions about chordate origins remains the evolution of the dorsal hollow nerve cord from the nervous system of a less centralized ancestor with little or no capacity for neurulation. In general, hypotheses imply that in the early embryo, the formation of neural ectoderm (prospective for motor neurons and interneurons) was increasingly repressed towards

one midline, and the neurulation process was induced along the edges of the narrowed neurectoderm territory. Thereafter Bmp exerted its neural patterning effects from the neural plate borders. This, of course, remains an area for future investigation[6,68,100].

Chordates differ from hemichordates in that Chordin and other Bmp antagonists are produced mostly in midline mesoderm, and specifically in mesoderm of Spemann's organizer, a region formed in the late blastula embryo at a location of high Nodal signalling and low Bmp signalling[101]. Organizer cells are precursors of the notochord and head mesoderm. Notochord precursors undergo extreme convergent extension by cell intercalation, forcefully repacking a cube of cells into a rod one-cell wide and lengthening the embryonic midline. Simultaneously they secrete their dorsoventral patterning molecules, neuralizing nearly half the embryo's overlying ectoderm and initiating neurulation morphogenesis. In chordates, neurogenic ectoderm produces little or no Bmp antagonist, whereas in hemichordates it is the main source. Recently, a possible notochord homologue, the axochord, was described for the polychaete annelid *P. dumerilii*; it is a midline mesodermal structure of muscle cells contained in a strong sheath into which lateral muscles attach, but there is no evidence that it is a notochord-like signalling source[102]. Rather, the midline signalling source of this protostome is presumably nearby neural ectoderm that determines, among other things, where the axochord itself develops.

Taken together, these data suggest that hemichordates are like protostomes in their dorsoventral development, whereas chordates have considerably modified the ancestral patterning mode (Fig. 4c), by adding organizer mesoderm as the Chordin source, and acquiring a large-scale neuralization response in the ectoderm. These innovations in DV axiation must be considered in any discussion of chordate origins.

Gill slits are a deuterostome innovation

The single unambiguous anatomical homology that is a clear deuterostome synapomorphy is the pharyngeal gill-slit complex[6,46,53,55,60,103]. These perforations of pharyngeal endoderm and ectoderm, ringed by beating cilia, imply that the ancestor fed by ingesting food particles carried by water flow entering the mouth and exiting the slits. The complex, which can include more than a hundred (bilaterally symmetrical) gill-slit pairs, is a major developmental and morphological modification beyond the bilaterian ancestor's pharynx, although presumably elaborated from it. Although hemichordates do not have a well-defined pharyngeal endostyle like chordates, the pharynx as a whole, and even the proboscis, probably makes endostyle-like mucociliary contributions to food trapping and conveyance to the gut[52,60]. Some of these functions may be deuterostome synapomorphies. To coordinate the functions of gill-slit-mediated water propulsion, food intake, trapping and conveyance, the pharyngeal nervous system is likely to have become modified from that of the bilateral ancestor. Given that pharyngeal innovations may represent the signature morphological, developmental and genomic innovations of deuterostomes, their development and physiology should be characterized more comprehensively.

Among extant animals, the filter-feeding lifestyle correlates with simplified body plans — radialized dorsoventral dimensions, more dispersed nervous systems, less cephalization of sensory systems, and less motility by trunk and tail axial muscles — when compared with extant food-seeking or predatory arthropods, annelids and jawed vertebrates. Such simplifications are presumably anatomical or physiological adaptations that benefit gill-slit-mediated filter feeding, and it seems plausible that evolution along the deuterostome stem involved considerable morphological modifications relative to earliest bilaterian body plans. If true, it is nonetheless apparent from AP and DV domain maps that the deuterostome ancestor suffered no concomitant loss of body-plan complexity at the molecular genetic developmental level. Rather, it shows that bilaterian domain maps are remarkably stable and can support wide-ranging morphologies, organogenesis and cytodifferentiations. An example of such modification is the muscular proboscis of hemichordates. The proboscis is used to dig and to trap food, while containing most of the conserved basic patterning elements of the

vertebrate forebrain, here spread over a basiepithelial nerve plexus[63,85]. There would be no intrinsic reason for the deuterostome ancestor to preserve the morphology and differentiations of the bilaterian ancestor if it no longer lived that ancestor's lifestyle. Finally, there is some palaeontological support for a filter-feeding deuterostome ancestor. On the basis of molecular clock estimates, deuterostome phyla would have diverged in the Ediacaran period, well before the Cambrian explosion. The lack of an obvious fossil record, except for small Precambrian trace fossils and the enigmatic Ediacaran fossils, and evidence of abundant filterable food sources in the form of microbial mats and plankton, suggest that bilaterians of that time were probably small and simple filter feeders[104–106].

This interpretation of the deuterostome ancestor has important consequences for the origin of chordates. Relative to that ancestor, the chordate stem lineage achieved major developmental and morphological innovations, including the evolution of a true notochord from the archenteron roof, centralizing many morphogenetic activities of the ancestral archenteron and taking over the signalling activities of the ancestral ectoderm for both AP and DV axial patterning (by producing Bmp and Wnt antagonists), to become the centrepiece of Spemann's organizer. Concomitantly, the innovations of neural induction (neuralization) and full-length neurulation of the ectoderm generated a hollow nerve cord along the entire body length (a length now defined by the elongating notochord), rather than just the short and late collar cord neurulation of hemichordates (although this limited neurulation shows that the ancestor possessed the basic morphogenetic process and components). At some point later, dorsoventral inversion of the chordate body took place, with mouth relocation out of the neural ectoderm[107]. All of this occurred on the chordate stem, perhaps after the elimination of an ambulacrarian-type larva, to open up uninterrupted embryonic development of the adult body plan (see ref. 100 for further elaboration of this hypothesis).

Future directions

Insights into deuterostome evolution are emerging from research in developmental biology, phylogenomics, genomics and zoology. A particular focus has been the pharyngeal gill-slit complex, which is supported as an ancestral deuterostome feature by strong morphological and developmental data. The implication that the deuterostome ancestor was a filter feeder naturally draws attention to other integrated pharyngeal specializations, including endostyle-like food-trapping organs. Further study of these organs, especially in amphioxus and hemichordates, has the immediate potential to reveal clues about deuterostome and chordate origins.

One of the most important differences between hemichordates and chordates, revealed by comparative developmental studies, is the source of Bmp antagonists involved in establishing DV axial polarity in early development. These antagonists are expressed in the ectoderm of hemichordates and the mesoderm of the chordate organizer. Spemann's organizer is a key chordate developmental innovation defined by various secreted factors modulating Bmp, Nodal and Wnt signalling. A more comprehensive description of the roles of these signals and their antagonists in the patterning of hemichordate mesoderm and ectoderm will be required to devise and test hypotheses about the evolution of the chordate organizer.

Most developmental insights from hemichordates have so far come from studies of direct developing hemichordates, but a distinct larval life-history stage is probably an ancestral trait of Ambulacraria and perhaps of deuterostomes (Box 2). More comprehensive developmental studies in indirect-developing echinoderms and hemichordates, with distinct larval body plans, are needed to determine the importance of complex life cycles and the role of larvae in the early diversification of deuterostome body plans. Garstang's influential auricularian hypothesis derived the chordate body plan from an ancestral larval body plan[108], but this hypothesis has recently lost support due to revisions in chordate phylogeny and close similarities between adult rather than larval body patterning[109]. Comparative data sets on larval patterning will be key for reconstructing ancestral developmental strategies of early

deuterostomes and testing hypotheses of larval homology.

Finally, advances in genomics have begun to shed light on the gene content and chromosomal organization of invertebrate deuterostomes, including the purple sea urchin (*Strongylocentrotus purpuratus*)[110], the acorn worms *S. kowalevskii* and *Ptychodera flava*[87,111], and the crown-of-thorns sea star (*Acanthaster planci*)[112]. Given the apparent conservation of not only the pan-deuterostome axial maps but also many of the downstream factors that control organogenesis, it will be exciting to explore the gene-regulatory elements that underlie this deep conservation through a combination of comparative genomics and experimental developmental biology, revealing features of the ancestral deuterostome down to the nucleotide level. ∎

1. Bateson, W. The ancestry of the chordata. *Q. J. Microsc. Sci.* **26,** 535–571 (1886).
2. Kowalevsky, A. Anatomie des Balanoglossus [in French]. *Mem. Acad. Imp. Sci. St Petersb.* **7,** 16 (1866).
3. Metchnikoff, V. E. Über die systematische Stellung von Balanoglossus [in German]. *Zool. Anz.* **4,** 139–157 (1881).
4. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* 2nd edn (Blackwell, 2005).
5. De Robertis, E. M. Evo-devo: variations on ancestral themes. *Cell* **132,** 185–195 (2008).
6. Brown, F. D., Prendergast, A. & Swalla, B. J. Man is but a worm: chordate origins. *Genesis* **46,** 605–613 (2008).
7. Gerhart, J., Lowe, C. & Kirschner, M. Hemichordates and the origin of chordates. *Curr. Opin. Genet. Dev.* **15,** 461–467 (2005).
8. Holland, L. Z. Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. *Nature Rev. Neurosci.* **10,** 736–746 (2009).
9. Lapraz, F., Besnardeau, L. & Lepage, T. Patterning of the dorsal-ventral axis in echinoderms: insights into the evolution of the BMP-chordin signaling network. *PLoS Biol.* **7,** e1000248 (2009).
10. Satoh, N. *Developmental Genomics of Ascidians* (Wiley-Blackwell, 2014).
11. Yu, J. K. *et al.* Axial patterning in cephalochordates and the evolution of the organizer. *Nature* **445,** 613–617 (2007).
    This article reports molecular evidence for a functional organizer in amphioxus, and supports the presence of this molecular patterning module at the base of the chordates.
12. Röttinger, E. & Lowe, C. J. Evolutionary crossroads in developmental biology: hemichordates. *Development* **139,** 2463–2475 (2012).
13. Angerer, L. M., Yaguchi, S., Angerer, R. C. & Burke, R. D. The evolution of nervous system patterning: insights from sea urchin development. *Development* **138,** 3613–3623 (2011).
14. Peter, I. S. & Davidson, E. H. A gene regulatory network controlling the embryonic specification of endoderm. *Nature* **474,** 635–639 (2011).
15. Swalla, B. J. & Smith, A. B. Deciphering deuterostome phylogeny: molecular, morphological and palaeontological perspectives. *Phil. Trans. R. Soc. Lond. B* **363,** 1557–1568 (2008).
16. Haeckel, E. H. P. A. *Anthropogenie: oder, Entwickelungsgeschichte des Menschen* [in German] (Wilhelm Engelmann, 1874).
17. Kowalevsky, A. Weitere Studien über die Entwicklung der einfachen Ascidien [in German]. *Arch. Mikr. Anat.* **7,** 101–130 (1866).
18. Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (D. Appleton and company, 1871).
19. Grobben, K. Die systematische einteilung des tierreiches [in German]. *Verh. der Zool.-Bot. Ges. Wien.* **58,** 491–511 (1908).
20. Martin-Duran, J. M., Janssen, R., Wennberg, S., Budd, G. E. & Hejnol, A. Deuterostomic development in the protostome *Priapulus caudatus*. *Curr. Biol.* **22,** 2161–2166 (2012).
21. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439,** 965–968 (2006).
    The authors of this paper provide conclusive molecular evidence for the sister relationship between tunicates and vertebrates and the basal position of cephalochordates.
22. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452,** 745–749 (2008).
23. Wada, H. & Satoh, N. Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18S rDNA. *Proc. Natl Acad. Sci. USA* **91,** 1801–1804 (1994).
24. Blair, J. E. & Hedges, S. B. Molecular phylogeny and divergence times of deuterostome animals. *Mol. Biol. Evol.* **22,** 2275–2284 (2005).
25. Field, K. G. *et al.* Molecular phylogeny of the animal kingdom. *Science* **239,** 748–753 (1988).
26. Halanych, K. M. The phylogenetic position of the pterobranch hemichordates based on 18S rDNA sequence data. *Mol. Phylogenet. Evol.* **4,** 72–76 (1995).
27. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* **470,** 255–258 (2011).
28. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. Biol. Sci.* **276,** 4261–4270 (2009).
29. Gee, H. *Before the Backbone* (Chapman & Hall, 1996).
30. Stach, T. Chordate phylogeny and evolution: a not so simple three-taxon problem. *J. Zool.* **276,** 117–141 (2008).
31. Lu, T. M., Luo, Y. J. & Yu, J. K. BMP and Delta/Notch signaling control the development of amphioxus epidermal sensory neurons: insights into the evolution of the peripheral sensory system. *Development* **139,** 2020–2030 (2012).
32. Escriva, H., Holland, N. D., Gronemeyer, H., Laudet, V. & Holland, L. Z. The retinoic acid signaling pathway regulates anterior/posterior patterning in the nerve cord and pharynx of amphioxus, a chordate lacking neural crest. *Development* **129,** 2905–2916 (2002).
33. Onai, T. *et al.* Retinoic acid and Wnt/β-catenin have complementary roles in anterior/posterior patterning embryos of the basal chordate amphioxus. *Dev. Biol.* **332,** 223–233 (2009).
34. Meulemans, D. & Bronner-Fraser, M. Gene-regulatory interactions in neural crest evolution and development. *Dev. Cell* **7,** 291–299 (2004).
35. Shimeld, S. M. The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. *Dev. Genes Evol.* **209,** 40–47 (1999).
36. Holland, L. Z. *et al.* Evolution of bilaterian central nervous systems: a single origin? *EvoDevo* **4,** 27 (2013).
37. Irimia, M. *et al.* Conserved developmental expression of *Fezf* in chordates and *Drosophila* and the origin of the *Zona Limitans Intrathalamica* (ZLI) brain organizer. *EvoDevo* **1,** 7 (2010).
38. Abedin, M. & King, N. The premetazoan ancestry of cadherins. *Science* **319,** 946–948 (2008).
39. Conklin, E. G. The embryology of amphioxus. *J. Morphol.* **54,** 69–151 (1932).
40. Gomez, C. *et al.* Control of segment number in vertebrate embryos. *Nature* **454,** 335–339 (2008).
41. Beaster-Jones, L. *et al.* Expression of somite segmentation genes in amphioxus: a clock without a wavefront? *Dev. Genes Evol.* **218,** 599–611 (2008).
42. Bertrand, S. *et al.* Amphioxus FGF signaling predicts the acquisition of vertebrate morphological traits. *Proc. Natl Acad. Sci. USA* **108,** 9160–9165 (2011).
43. Kozmik, Z. *et al.* Pax-Six-Eya-Dach network during amphioxus development: conservation *in vitro* but context specificity *in vivo*. *Dev. Biol.* **306,** 143–159 (2007).
44. Mazet, F., Amemiya, C. T. & Shimeld, S. M. An ancient Fox gene cluster in bilaterian animals. *Curr. Biol.* **16,** R314–316 (2006).
45. Jandzik, D. *et al.* Evolution of the new vertebrate head by co-option of an ancient chordate skeletal tissue. *Nature* **518,** 534–537 (2015).
46. Rychel, A. L. & Swalla, B. J. Development and evolution of chordate cartilage. *J. Exp. Zool. B Mol. Dev. Evol.* **308,** 325–335 (2007).
47. Wright, G. M., Keeley, F. W. & Robson, P. The unusual cartilaginous tissues of jawless craniates, cephalochordates and invertebrates. *Cell Tissue Res.* **304,** 165–174 (2001).
48. Range, R. C., Angerer, R. C. & Angerer, L. M. Integration of canonical and noncanonical Wnt signaling pathways patterns the neuroectoderm along the anterior–posterior axis of sea urchin embryos. *PLoS Biol.* **11,** e1001467 (2013).
49. Yankura, K. A., Martik, M. L., Jennings, C. K. & Hinman, V. F. Uncoupling of complex regulatory patterning during evolution of larval development in echinoderms. *BMC Biol.* **8,** 143 (2010).
50. Sato, A., Bishop, J. D. & Holland, P. W. Developmental biology of pterobranch hemichordates: history and perspectives. *Genesis* **46,** 587–591 (2008).
51. Dominguez, P., Jacobson, A. G. & Jefferies, R. P. Paired gill slits in a fossil with a calcite skeleton. *Nature* **417,** 841–844 (2002).
52. Gonzalez, P. & Cameron, C. B. The gill slits and pre-oral ciliary organ of *Protoglossus* (Hemichordata: Enteropneusta) are filter-feeding structures. *Biol. J. Linn. Soc.* **98,** 898–906 (2009).
53. Gillis, J. A., Fritzenwanker, J. H. & Lowe, C. J. A stem-deuterostome origin of the vertebrate pharyngeal transcriptional network. *Proc. Biol. Sci.* **279,** 237–246 (2012).
    This report shows extensive patterning similarities between the development of chordate and enteropneust gills, further supporting morphological homology.
54. Ogasawara, M., Wada, H., Peters, H. & Satoh, N. Developmental expression of *Pax1/9* genes in urochordate and hemichordate gills: insight into function and evolution of the pharyngeal epithelium. *Development* **126,** 2539–2550 (1999).
55. Fritzenwanker, J. H., Gerhart, J., Freeman, R. M. Jr & Lowe, C. J. The Fox/Forkhead transcription factor family of the hemichordate *Saccoglossus kowalevskii*. *EvoDevo* **5,** 17 (2014).
56. Balser, E. J. & Ruppert, E. E. Structure, ultrastructure, and function of the preoral heart-kidney in *Saccoglossus kowalevskii* (Hemichordata, Enteropneusta) including new data on the stomochord. *Acta Zool.* **71,** 235–249 (1990).
57. Miyamoto, N. & Wada, H. Hemichordate neurulation and the origin of the neural tube. *Nature Commun.* **4,** 2713 (2013).
    This manuscript demonstrates similar mediolateral patterning mechanisms between the hemichordate collar cord and chordate dorsal cord.
58. Hyman, L. H. *The Invertebrates* 1st edn (McGraw-Hill, 1959).
59. Luttrell, S., Konikoff, C., Byrne, A., Bengtsson, B. & Swalla, B. J. Ptychoderid hemichordate neurulation without a notochord. *Integr. Comp. Biol.* **52,** 829–834 (2012).
60. Ruppert, E. E. Key characters uniting hemichordates and chordates: homologies or homoplasies? *Can. J. Zool.* **83,** 8–23 (2005).
61. Satoh, N. *et al.* On a possible evolutionary link of the stomochord of hemichordates to pharyngeal organs of chordates. *Genesis* **52,** 925–934 (2014).
62. De Robertis, E. M. & Kuroda, H. Dorso-ventral patterning and neural induction in *Xenopus* embryos. *Annu. Rev. Cell Dev. Biol.* **20,** 285–308 (2004).

63. Pani, A. M. *et al*. Ancient deuterostome origins of vertebrate brain signalling centres. *Nature* **483**, 289–294 (2012).
    This paper presents evidence that ectodermal signalling centres thought to have been uniquely associated with the evolution of vertebrate brains are present in hemichordates as part of a conserved ancient deuterostome patterning network.
64. Bullock, T. H. The anatomical organization of the nervous system of enteropneusta. *Q. J. Microsc. Sci.* **86**, 55–111 (1945).
65. Kaul, S. & Stach, T. Ontogeny of the collar cord: neurulation in the hemichordate *Saccoglossus kowalevskii*. *J. Morphol.* **271**, 1240–1259 (2010).
66. Knight-Jones, E. On the nervous system of *Saccoglossus cambriensis* (Enteropneusta). *Phil. Trans. R. Soc. Lond. B* **236**, 315–354 (1952).
67. Morgan, T. Development of *Balanoglossus*. *J. Morphol.* **9**, 1–86 (1894).
68. Benito-Gutierrez, E. & Arendt, D. CNS evolution: new insight from the mud. *Curr. Biol.* **19**, R640–642 (2009).
69. Nomaksteinsky, M. *et al*. Centralization of the deuterostome nervous system predates chordates. *Curr. Biol.* **19**, 1264–1269 (2009).
    This paper shows clear molecular evidence for the presence of cell bodies in the dorsal nerve cord of enteropneusts and proposes the deep ancestry of a CNS in the deuterostomes.
70. Bullock, T. H. The functional organisation of the nervous system of the Enteropneusta. *Biol. Bull.* **79**, 91–113 (1940).
71. Cameron, C. B. & Mackie, G. O. Conduction pathways in the nervous system of *Saccoglossus* sp. (Enteropneusta). *Can. J. Zool.* **74**, 15–19 (1996).
72. Lowe, C. J. *et al*. Dorsoventral patterning in hemichordates: insights into early chordate evolution. *PLoS Biol.* **4**, e291 (2006).
    This manuscript demonstrates the role of BMP signalling in the formation of the DV axis of the enteropneust adult body plan.
73. Martindale, M. Q. & Hejnol, A. A developmental perspective: changes in the position of the blastopore during bilaterian evolution. *Dev. Cell* **17**, 162–174 (2009).
74. Slack, J. M., Holland, P. W. & Graham, C. F. The zootype and the phylotypic stage. *Nature* **361**, 490–492 (1993).
75. Darras, S., Gerhart, J., Terasaki, M., Kirschner, M. & Lowe, C. J. β-Catenin specifies the endomesoderm and defines the posterior organizer of the hemichordate *Saccoglossus kowalevskii*. *Development* **138**, 959–970 (2011).
76. Wikramanayake, A. H., Huang, L. & Klein, W. H. β-Catenin is essential for patterning the maternally specified animal-vegetal axis in the sea urchin embryo. *Proc. Natl Acad. Sci. USA* **95**, 9343–9348 (1998).
77. Henry, J. Q., Perry, K. J., Wever, J., Seaver, E. & Martindale, M. Q. β-Catenin is required for the establishment of vegetal embryonic fates in the nemertean, *Cerebratulus lacteus*. *Dev. Biol.* **317**, 368–379 (2008).
78. Wikramanayake, A. H. *et al*. An ancient role for nuclear β-catenin in the evolution of axial polarity and germ layer segregation. *Nature* **426**, 446–450 (2003).
79. Kimelman, D. Mesoderm induction: from caps to chips. *Nature Rev. Genet.* **7**, 360–372 (2006).
80. Green, S. A., Norris, R. P., Terasaki, M. & Lowe, C. J. FGF signaling induces mesoderm in the hemichordate *Saccoglossus kowalevskii*. *Development* **140**, 1024–1033 (2013).
81. Hinman, V. F. & Davidson, E. H. Evolutionary plasticity of developmental gene regulatory network architecture. *Proc. Natl Acad. Sci. USA* **104**, 19404–19409 (2007).
82. Hikasa, H. & Sokol, S. Y. Wnt signaling in vertebrate axis specification. *Cold Spring Harb. Perspect. Biol.* **5**, a007955 (2013).
83. Petersen, C. P. & Reddien, P. W. Wnt signaling and the polarity of the primary body axis. *Cell* **139**, 1056–1068 (2009).
84. Kiecker, C. & Niehrs, C. A morphogen gradient of Wnt/β-catenin signalling regulates anteroposterior neural patterning in *Xenopus*. *Development* **128**, 4189–4201 (2001).
85. Lowe, C. J. *et al*. Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* **113**, 853–865 (2003).
    This paper provides evidence for a conserved transcriptional gene regulatory network between hemichordates and chordates despite large organizational differences in their basic body plans.
86. Aronowicz, J. & Lowe, C. J. Hox gene expression in the hemichordate *Saccoglossus kowalevskii* and the evolution of deuterostome nervous systems. *Integr. Comp. Biol.* **46**, 890–901 (2006).
87. Freeman, R. *et al*. Identical genomic organization of two hemichordate hox clusters. *Curr. Biol.* **22**, 2053–2058 (2012).
88. Holland, L. Z. *et al*. Evolution of bilaterian central nervous systems: a single origin? *EvoDevo* **4**, 27 (2013).
89. David, B. & Mooi, R. How Hox genes can shed light on the place of echinoderms among the deuterostomes. *EvoDevo* **5**, 22 (2014).
90. Hara, Y. *et al*. Expression patterns of Hox genes in larvae of the sea lily *Metacrinus rotundus*. *Dev. Genes Evol.* **216**, 797–809 (2006).
91. Morris, V. B. & Byrne, M. Oral-aboral identity displayed in the expression of HpHox3 and HpHox11/13 in the adult rudiment of the sea urchin *Holopneustes purpurescens*. *Dev. Genes Evol.* **224**, 1–11 (2014).
92. Lacalli, T. Echinoderm conundrums: Hox genes, heterochrony, and an excess of mouths. *Evodevo* **5**, 46 (2014).
93. Peterson, K. J., Arenas-Mena, C. & Davidson, E. H. The A/P axis in echinoderm ontogeny and evolution: evidence from fossils and molecules. *Evol. Dev.* **2**, 93–101 (2000).
94. Omori, A., Akasaka, K., Kurokawa, D. & Amemiya, S. Gene expression analysis of Six3, Pax6, and Otx in the early development of the stalked crinoid *Metacrinus rotundus*. *Gene Expr. Patterns* **11**, 48–56 (2011).
95. Arenas-Mena, C., Cameron, A. R. & Davidson, E. H. Spatial expression of Hox cluster genes in the ontogeny of a sea urchin. *Development* **127**, 4631–4643 (2000).
96. De Robertis, E. M. & Sasai, Y. A common plan for dorsoventral patterning in Bilateria. *Nature* **380**, 37–40 (1996).
97. Mizutani, C. M. & Bier, E. EvoD/Vo: the origins of BMP signalling in the neuroectoderm. *Nature Rev. Genet.* **9**, 663–677 (2008).
98. Denes, A. S. *et al*. Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell* **129**, 277–288 (2007).
99. Röttinger, E. & Martindale, M. Q. Ventralization of an indirect developing hemichordate by NiCl₂ suggests a conserved mechanism of dorso-ventral (D/V) patterning in Ambulacraria (hemichordates and echinoderms). *Dev. Biol.* **354**, 173–190 (2011).
100. Satoh, N. An aboral-dorsalization hypothesis for chordate origin. *Genesis* **46**, 614–622 (2008).
101. Gerhart, J. Evolution of the organizer and the chordate body plan. *Int. J. Dev. Biol.* **45**, 133–153 (2001).
102. Lauri, A. *et al*. Development of the annelid axochord: insights into notochord evolution. *Science* **345**, 1365–1368 (2014).
103. Cameron, C. B., Garey, J. R. & Swalla, B. J. Evolution of the chordate body plan: new insights from phylogenetic analyses of deuterostome phyla. *Proc. Natl Acad. Sci. USA* **97**, 4469–4474 (2000).
    This manuscript proposes that the extant enteropneust, rather than pterobranch, adult body plan may best represent ancestral deuterostome characters.
104. Erwin, D. H. *et al*. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* **334**, 1091–1097 (2011).
105. Erwin, D. H. & Valentine, J. *The Cambrian Explosion: the Construction of Animal Biodiversity* (Roberts and Company, 2013).
106. Peterson, K. J. & Butterfield, N. J. Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proc. Natl Acad. Sci. USA* **102**, 9547–9552 (2005).
107. Lacalli, T. C. The emergence of the chordate body plan: some puzzles and problems. *Acta Zool.* **91**, 4–10 (2010).
108. Garstang, W. The morphology of the Tunicata. *Q. J. Microsc. Sci.* **72**, 51–187 (1928).
109. Lacalli, T. C. Protochordate body plan and the evolutionary role of larvae: old controversies resolved? *Can. J. Zool.* **83**, 216–224 (2005).
110. Sea Urchin Genome Sequencing Consortium. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941–952 (2006).
111. Freeman, R. M. Jr *et al*. cDNA sequences for transcription factors and signaling proteins of the hemichordate *Saccoglossus kowalevskii*: efficacy of the expressed sequence tag (EST) approach for evolutionary and developmental studies of a new organism. *Biol. Bull.* **214**, 284–302 (2008).
112. Baughman, K. W. *et al*. Genomic organization of Hox and ParaHox clusters in the echinoderm, *Acanthaster planci*. *Genesis* **52**, 952–958 (2014).
113. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968 (2006).
114. Nielsen, C. Origin of the chordate central nervous system — and the origin of chordates. *Dev. Genes Evol.* **209**, 198–205 (1999).
115. Gudo, M. & Syed, T. 100 years of Deuterostomia (Grobben, 1908): Cladogenetic and anagenetic relations within the notoneuralia domain. http://arxiv.org/abs/0811.2189 (2008).
116. Jefferies, R. P. S. *The Ancestry of the Vertebrates* (Cambridge Univ. Press, 1986).
117. Swalla, B. J. Building divergent body plans with similar genetic pathways. *Heredity* **97**, 235–243 (2006).
118. Bone, Q. The central nervous system in amphioxus. *J. Comp. Neurol.* **115**, 27–51 (1960).
119. Wicht, H. & Lacalli, T. C. The nervous system of amphioxus: structure, development, and evolutionary significance. *Can. J. Zool.* **83**, 122–150 (2005).
120. Cannon, J. T. *et al*. Phylogenomic resolution of the hemichordate and echinoderm clade. *Current Biol.* **24**, 2827–2832 (2014).
121. Cameron, C. B. Particle retention and flow in the pharynx of the enteropneust worm *Harrimania planktophilus*: the filter-feeding pharynx may have evolved before the chordates. *Biol. Bull.* **202**, 192–200 (2002).
122. Telford, M. J. *et al*. Phylogenomic analysis of echinoderm class relationships supports Asterozoa. *Proc. Biol. Sci.* **281**, 20140479 (2014).

REVIEW

A new heart for a new head in vertebrate cardiopharyngeal evolution

Rui Diogo[1], Robert G. Kelly[2], Lionel Christiaen[3], Michael Levine[4], Janine M. Ziermann[1], Julia L. Molnar[1], Drew M. Noden[5] & Eldad Tzahor[6]

It has been more than 30 years since the publication of the new head hypothesis, which proposed that the vertebrate head is an evolutionary novelty resulting from the emergence of neural crest and cranial placodes. Neural crest generates the skull and associated connective tissues, whereas placodes produce sensory organs. However, neither crest nor placodes produce head muscles, which are a crucial component of the complex vertebrate head. We discuss emerging evidence for a surprising link between the evolution of head muscles and chambered hearts — both systems arise from a common pool of mesoderm progenitor cells within the cardiopharyngeal field of vertebrate embryos. We consider the origin of this field in non–vertebrate chordates and its evolution in vertebrates.

In their influential 1983 paper, Gans and Northcutt[1] proposed that early vertebrates evolved from invertebrates principally through innovations in the head. These include the muscularization of the ventrolateral mesoderm, or hypomere, to form branchiomeric muscles and the emergence of two novel ectodermal structures: the neurogenic placodes and the neural crest. Neural crest cells produce most of the cartilage, bone, dentine and other connective tissues of the vertebrate head, whereas the placodes give rise to the sensory neurons that are essential for the formation of vertebrates' complex sensory systems[2–4]. The new head hypothesis proposed that these evolutionary innovations were associated with a shift from passive filter-feeding to active predation. Increased sensory capabilities and a muscularized pharynx arguably permitted more efficient prey detection and capture, as well as higher rates of respiratory gas exchange, which accompany the predatory lifestyle. This major behavioural and ecological transition also coincided with the emergence of a chambered heart, which presumably allowed for the increased growth and metabolism that was demanded by active predation. However, the new head hypothesis was primarily concerned with derivatives of neural crest and placodes, which are better represented in the fossil record than soft tissues such as muscles[5,6]. In this Review, we provide an up-to-date multidisciplinary discussion of the origin and evolution of vertebrate head muscles, taking into account surprising new evidence for shared developmental origins of several head muscles and the heart, and the ancient (pre-vertebrate) origin of this association.

The emerging concept of the cardiopharyngeal field

The cardiopharyngeal field (CPF) is a developmental domain that gives rise to the heart and branchiomeric muscles (Box 1 and Figs 1, 2). The amniote heart is made up of cardiomyocytes derived from two adjacent progenitor cell populations in the early embryo[7]. Early differentiating cardiac progenitor cells of the first heart field (FHF) give rise to the linear heart tube and later form the left ventricle and parts of the atria[8,9]. Subsequently, second-heart-field (SHF) progenitors, located in pharyngeal mesoderm, produce cardiac muscle tissue (myocardium) of the outflow tract, right ventricle and parts of the atria[10–12] (Fig. 2). The SHF can be divided into anterior and posterior progenitor cell populations that contribute to the arterial and venous poles of the heart, respectively[8]. Cells from pharyngeal mesoderm can form either cardiac or skeletal muscles, depending on signals from adjacent pharyngeal endoderm, surface ectoderm and neural crest cells[9,13–16]. The latter have important roles in regulating the development of the CPF — they are required for the deployment of SHF-derived cells to the heart's arterial pole, and neural-crest-derived mesenchyme patterns branchiomeric muscle formation and gives rise to associated fascia and tendons[17–19].

A suite of regulatory factors integrates the intercellular signals that coordinate the formation of cardiac and branchiomeric muscles from a common pool of mesodermal progenitor cells. Within the CPF there is considerable overlap in the expression of genes that encode cardiogenic regulatory factors (for example, *Isl1* (also known as *Islet1*) and *Nkx2-5*) and those that specify head muscles (for example, *Tbx1*, *Tcf21* (also known as capsulin), *Msc* (also known as *MyoR*) and *Pitx2*)[13,15,20]. Importantly, many of the intercellular signalling pathways and transcription factors that control branchiomeric myogenesis upstream of the MyoD family of myogenic determination factors differ fundamentally from those operating in the trunk[21,22]. Here we focus on Isl1, Nkx2-5 and Tbx1. The LIM-homeodomain protein Isl1 is required in a broad subset of cardiovascular progenitor cells in mouse embryos[23] and it is expressed in pharyngeal mesoderm, including the pharyngeal arches and SHF. Isl1+ progenitor cells substantially contribute to the heart and branchiomeric muscles, but not to hypobranchial (for example, tongue) or extraocular (eye) muscles[13,24]. Expression and functional studies indicate that Isl1 delays differentiation of branchiomeric muscles[13,24]; Isl1 thus marks a subset of CPF cells and plays an important part in the development of distinct cardiovascular and skeletal muscle progenitors[24]. The cardiac transcription factor Nkx2-5 regulates proliferation in the SHF and acts with Isl1 to modulate SHF progenitor-specific gene expression[25–27]. *Tbx1* is required within the CPF for both heart and head muscle development, and is the major candidate gene for the congenital condition DiGeorge syndrome (or 22q11.2 deletion syndrome), which is characterized by a spectrum of cardiovascular defects and craniofacial anomalies. Like Isl1, Tbx1 has a crucial and conserved role in extending the heart's arterial pole by promoting proliferation and delaying differentiation of SHF cells[28–31]. Tbx1 is also required for activation of branchiomeric myogenesis and may directly regulate the myogenic determination gene *MyoD*[32–34].

[1]Department of Anatomy, Howard University College of Medicine, Washington DC 20059, USA. [2]Aix Marseille Université, Centre National de la Recherche Scientifique, Institut de Biologie du Développement de Marseille UMR 7288, 13288 Marseille, France. [3]Center for Developmental Genetics, Department of Biology, New York University, New York 10003, USA. [4]Department of Molecular and Cell Biology, University of California at Berkeley, California 94720, USA. [5]Department of Biomedical Sciences, College of Veterinary Medicine , Cornell University, Ithaca, New York 14853, USA. [6]Department of Biological Regulation, Weizmann Institute of Science, Rehovot 76100, Israel.

Tbx1 acts upstream of the LIM-homeodomain protein Lhx2 within an intricate regulatory network that specifies cardiopharyngeal progenitors. Genetic ablation of these factors, alone or in combination, results in cardiac and head muscle defects; including DiGeorge syndrome phenotypes[35]. Thus, evolutionarily conserved regulatory factors maintain a pool of cardiopharyngeal progenitor cells for SHF-specific cardiogenesis and branchiomeric myogenesis.

Confirmation that multipotent progenitor cells give rise to branchiomeric skeletal muscles and SHF-derived regions of the heart comes from retrospective clonal analyses in mice, a method for analysing cell lineage in intact embryos[36]. These experiments demonstrated the existence of a series of common cardiopharyngeal progenitors along the anteroposterior axis that contribute to heart-tube growth and branchiomeric muscle morphogenesis. Interestingly, comparative anatomists suggested decades ago that branchiomeric muscles are related to muscles derived from the 'visceral' mesoderm (for example, of the heart and anterior gut)[37,38], a view supported by the recent genetic and developmental studies reviewed here. Moreover, mouse clonal analyses revealed relationships between specific regions of the heart and subsets of branchiomeric muscles that go beyond the predictions of early comparative anatomists. SHF-derived regions of the heart, for example, are developmentally more closely related to branchiomeric muscles than to FHF-derived regions of the heart[7,36]. In support of such a grouping, the cardiac lineages contributing to the FHF and SHF have been shown to diverge before expression of *Mesp1* during early gastrulation[39,40]. Taken together, recent findings provide a new paradigm for exploring the collinear emergence of cardiac chambers and branchiomeric muscles that underlies the early evolution and diverse origins of the vertebrate head[9,21,22,41,42].

Origins and diversity of cardiopharyngeal structures

The heads of mammals, including humans, contain more than 60 muscles[43], which control eye movements and allow food uptake, respiration, and facial and vocal communication[44–46]. Strikingly, the human head includes at least six different groups of muscles with distinct developmental origins and evolutionary histories[35,37,44] (Fig. 1). Full recognition and detailed knowledge of this heterogeneity has enormous basic science and clinical implications because long accepted anatomy concepts, mainly based on adult function and physiology (for example, skeletal compared with cardiac muscles) do not correspond to the true developmental and evolutionary origins of body structures. Even the conventional classification of head muscle groups based on topographical relations masks the true heterogeneity of muscle origins and progenitor fates (for example, molecular profiling of early determinative signalling molecules and transcription factors reveals almost as much heterogeneity within each group — such as, branchial, extraocular and tongue — as between them[43]).

Comparative anatomical studies identified homologues of many amniote branchiomeric muscles in gnathostome (jawed) fish such as sharks, suggesting that they have ancient origins[47,48] (Fig. 3). Cyclostomes (hagfish and lampreys[49–52]) lack some of these muscles (for example, the cucullaris group), but like some chondrichthyans (Selachii and Holocephali) they possess an additional, seventh group of head muscles: epibranchial muscles, which are derived from anterior somites[53]. Thus, extraocular, branchiomeric, and both hypobranchial and epibranchial somite-derived muscles were integral parts of the heterogenous head musculature of early vertebrates[54–57] (Fig. 3). Moreover, lamprey embryos express homologues of *Isl1*, *Nkx2-5* and *Tbx1* in seemingly overlapping anterior and ventral mesodermal domains[58–61], comparable with the patterns of their homologues in the amniote CPF. Interestingly, the emergence of heterogeneous head-muscle groups at the base of vertebrates coincided with the emergence of chambered hearts[62,63] (Fig. 3). This intriguing correlation suggests that the two innovations are linked by their common developmental origin in the CPF.

Studies indicate that specific branchiomeric muscles were crucial for evolutionary innovations among vertebrates, such as the emergence of the tetrapod neck. The amniote neck muscles trapezius and sternocleidomastoideus (Fig. 1) derive from the cucullaris, a muscle

that probably appeared in early gnathostomes and was found in fossil placoderms[5,6,48,64,65]. Among extant gnathostomes, some of the anatomical and developmental characteristics of the cucullaris are shared with branchiomeric and somite-derived limb, epibranchial and hypobranchial muscles[57,66,67]. Most available data, however, indicate that the cucullaris is a branchiomeric muscle derived from the posterior-most pharyngeal arches, as suggested by Edgeworth[22,68–71]. Like other branchiomeric muscles, in most gnathostomes the cucullaris is attached to neural-crest-derived tendinous and skeletal elements[38,64,65,70,72]. Furthermore, *Tbx1* is active in core branchiomeric muscles (for example, the first and second arch muscles) and in the cucullaris-derived trapezius, whereas *Pax3* is required in the somites for limb, diaphragm, tongue, infrahyoid and trunk-muscle formation, but not for trapezius formation[22,73]. These findings may also support Gegenbaur's hypothesis that the pectoral appendage, to which the cucullaris and its derivatives usually attach, probably originated as an integral part of the head[74,75]. Thus, the evolutionary history of the cucullaris-related muscles illustrates the roles that branchiomeric muscles had in fostering anatomical and functional innovations during vertebrate evolution. Future studies are needed to investigate whether the emergence of the cucullaris at the base of gnathostomes coincided with cardiovascular innovations and, if so, whether this muscle also shares a common origin with a specific heart region (Fig. 1).

A urochordate cardiopharyngeal ontogenetic motif

Recent phylogenetic studies place the urochordates — not the cephalochordates (for example, amphioxus) — as the sister group of the
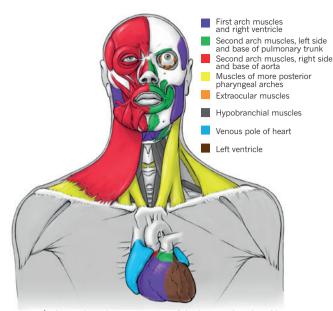
Figure 1 | The striking heterogeneity of the human head and heart musculature. The head includes at least six different muscle groups, all arising from the cardiopharyngeal field and being branchiomeric, except the hypobranchial and perhaps the extraocular muscles. On the left side of the body (right part of figure) the facial expression muscles have been removed to show the masticatory muscles. The six groups are: first/mandibular arch muscles, including cells clonally related to the right ventricle; left second/hyoid arch muscles related to myocardium at the base of the pulmonary trunk; right second/hyoid arch muscles, related to myocardium at the base of the aorta; muscles of the most posterior pharyngeal arches, including muscles of the pharynx and larynx and the cucullaris-derived neck muscles trapezius and sternocleidomastoideus; extraocular muscles, which are often not considered to be branchiomeric, but according to classic embryological studies and recent retrospective clonal analyses in mice contain cells related to those of the branchiomeric mandibular muscles; and hypobranchial muscles, including tongue and infrahyoid muscles that derive from somites and migrate into the head and neck[36,38,70].

vertebrates[76,77]. On the basis of these results, urochordates provide important insights for our understanding of the origin of vertebrates' evolutionary innovations, particularly from molecular and developmental perspectives. For instance, the new head hypothesis proposed that the emergence of branchiomeric muscles occurred during the transitions that led to the origin of vertebrates, and was associated with a shift from 'passive' filtration to more active feeding modes[1,4,78,79] and the emergence of crest- and placode-derived sensory organs. However, recent studies have identified neural-crest-like cells, placodes and a CPF in tadpole-like larvae of the ascidian *Ciona intestinalis*, a model urochordate (Figs 2, 4). The pan-placodal regulatory gene *Six1/2* is expressed in a crescent of cells straddling the anterior-most region of the developing neural tube in *C. intestinalis* embryos, comparable with the sites of origin of cranial placodes in the fate maps of vertebrates[80–82]. Ectodermal thickenings derived from this domain express placodal regulatory genes, including *Six3/6*, *Pitx* and *Eya*. For example, the atrial siphon placode shares extensive similarities with the vertebrate otic placode[3,80,81] (Fig. 4), whereas the stomodeum (the oral siphon primordium) expresses regulatory genes implicated in the specification of the vertebrate olfactory and adenohypophyseal placodes, including *Six*, *Eya* and the anterior placode markers *Pitx*[83–85] and *Dlx*. These new findings argue for homologies between urochordate siphon primordia and vertebrate placodes and suggest that; although certain placodes (profundal, maxillomandibular, epibranchial and lens) evolved by diversification within the vertebrate lineage[3], others (adenohypophyseal, olfactory and otic) appeared before the separation of vertebrates and urochordates (Figs 3, 4).

Ascidians and other urochordates possess a surprisingly sophisticated beating heart (Figs 2, 4), which shares several features with vertebrate

hearts, including localized pacemakers that drive a regular, rhythmic beat. The ascidian heart is derived from two Mesp+ cells in early embryos. These produce four trunk ventral cells, which express homologues of *Nkx2-5*, *Gata4, 5* and *6* and *Hand*, and migrate towards the pharyngeal endoderm[86–92]. They subsequently divide asymmetrically to produce medial heart precursors and secondary trunk ventral cells that divide again to produce second heart precursors and atrial siphon muscle precursors, which migrate towards the atrial siphon placode[93–95] (Figs 2, 4). Thus, trunk ventral cells are multipotent cardiopharyngeal progenitors that produce bona fide heart and pharyngeal muscles, following a clonal pattern evocative of that seen in mice (Fig. 2). Gene-expression profiling data are also consistent with the idea that the trunk ventral cells are homologous to the vertebrate cardiopharyngeal progenitors: trunk ventral cells express *Nk4*, the homologue of *Nkx2-5*, and secondary trunk ventral cells also express *Tbx1/10*, which is active in vertebrate pharyngeal mesoderm. Furthermore, the regulatory network governing interactions among the cardiopharyngeal specification genes seems to be highly conserved in ascidians and vertebrates. For example, cross-repressive interactions between Tbx1/10 and Nk4/Nkx2-5 delineate atrial siphon muscles and heart, respectively[95]. *Isl* is also expressed in the CPF, although there are differences from the precise expression profile seen in vertebrates, where Isl1 is thought to delay muscle differentiation[24]. It is nonetheless striking that all of the identified molecular determinants of the vertebrate SHF are expressed in ascidian trunk ventral cells.

There are additional parallels between the CPFs of ascidians and vertebrates in the regulatory circuitry underlying the differentiation of specialized muscles (Fig. 2). COE/Ebf functions downstream of Tbx1/10 and upstream of both Mrf/MyoD and Notch signalling to promote either early muscle differentiation or maintain undifferentiated precursors that produce most later atrial siphon and longitudinal muscles[93,96] (Fig. 2). Atrial siphon muscle precursors also associate with the Dlx+ atrial siphon placodes to form a ring of cells underlying the rosette-shaped placode in *C. intestinalis* swimming larvae[80,81,93,97]. These events parallel the migration of vertebrate branchiomeric muscle precursors into pharyngeal arches, their association with Dlx+ cranial neural crest cells, and the maintenance and growth of a pool of undifferentiated progenitor cells[24,98]. It is noteworthy that the ascidian FHF and SHF are each initially composed of four cells that independently arise from one of four multipotent cardiopharyngeal progenitors following a sequence of conserved regulatory interactions onto a stereotyped clonal pattern, producing FHF precursors and more closely related SHF and pharyngeal muscle precursors[95]. We refer to this clonal sequence of cell divisions, gene expression and cell-fate choices as a cardiopharyngeal ontogenetic motif[95] (Fig. 2).

Chordate origins of branchiomeric muscles

Studies using cephalochordates further probed the early chordate origins of branchiomeric-like pharyngeal muscles (Figs 3, 4). In the cephalochordate amphioxus, the larval mouth and unpaired primary gills develop five groups of orobranchial muscles[99,100]. This musculature is anatomically reminiscent of the vertebrate branchiomeric muscles, and disappears through apoptosis during metamorphosis to give way to adult oral, velar and pterygial muscles[99] (Fig. 4), which are even more similar to vertebrate adult branchiomeric muscles. The oral and velar muscles, in particular, share anatomical similarities with the oral and velar muscles of lampreys and hagfish (Fig. 4), although the pterygial muscles have a branchiomeric-like innervation pattern[99]. Gans[79] recognized this latter point and noted that this could mean that the branchiomeric muscles evolved before the last common ancestor (LCA) of vertebrates, as suggested by earlier authors[22], but contrary to the original new head hypothesis[1]. Vestigial muscles appear transiently with secondary gill formation in amphioxus, providing additional evidence that bilateral muscular gills and a segmental pattern of branchiomeric muscles were already present in the LCA of extant chordates[22].

Molecular studies suggest that the amphioxus homologues of *Tbx1*, *Nkx2-5* and *Isl1* are expressed in overlapping mesodermal domains in the pharyngeal region[101–103]. This domain includes cells that also express the
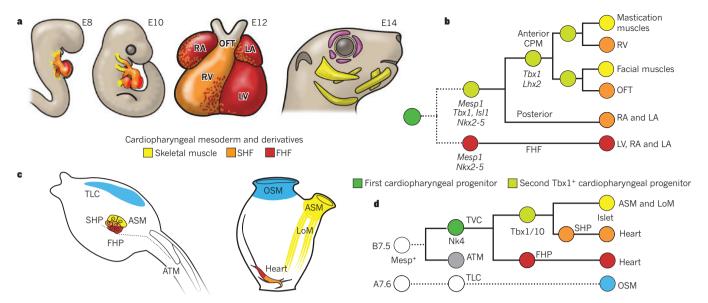
Figure 2 | An evolutionarily conserved cardiopharyngeal ontogenetic motif. a, Mouse embryos at embryonic days (E)8 and 10, the four-chambered mouse heart at E12, and the mouse head at E14. First heart field (FHF)-derived regions of heart (left ventricle (LV) and atria) are in red; second heart field (SHF)-derived regions of heart (right ventricle (RV), left atrium (LA), right atrium (RA) and outflow tract (OFT)) are in orange; branchiomeric skeletal muscles are in yellow; extraocular muscles are in purple. **b**, Lineage tree depicting the origins of cardiac compartments and branchiomeric muscles in mice. All cells derive from common pan-cardiopharyngeal progenitors (dark green) that produce the FHF, precursors of the left ventricle and atria, and the second Tbx1[+] cardiopharyngeal progenitors (light green). Broken lines indicate that the early common FHF and SHF progenitor remains to be identified in mice. In anterior cardiopharyngeal mesoderm (CPM), progenitor cells activate Lhx2, self-renew and produce the SHF-derived RV and OFT, and first and second arch branchiomeric muscles (including muscles of mastication and facial expression). **c**, Cardiopharyngeal precursors in *Ciona intestinalis* hatching

larva (left) and their derivatives in the metamorphosed juvenile (right). The first heart precursors (FHP) (red) and second heart precursors (SHP) (orange) contribute to the heart (red and orange mix), whereas atrial siphon muscle precursors (ASM, yellow) form atrial siphon and longitudinal muscles (LoM, yellow). Oral siphon muscles (OSM, blue) derive from a heterogenous larval population of trunk lateral cells (TLC, blue). ATM, anterior tail muscles. CPM is bilaterally symmetrical around the midline (dotted line). **d**, Lineage tree depicting clonal relationships and gene activities deployed in *C. intestinalis* cardiopharyngeal precursors. All cells derive from Mesp[+] B7.5 blastomeres, which produce ATM (grey, see also left panel of **c**) and trunk ventral cells (TVC, dark green). The latter pan-cardiopharyngeal progenitors express Nk4 and divide asymmetrically to produce the FHP (red) and second TVCs, the Tbx1/10[+] second cardiopharyngeal progenitors (second TVC, light green disk). The latter divide again asymmetrically to produce SHP (orange) and the precursors of ASM and LoM, which upregulate Islet. The OSM arise from A7.6-derived trunk lateral cells (TLC, light blue).

vertebrate cardiac markers *Hand* and *Tbx20* (refs 59, 104) and is thought to produce the branchial artery, a possible — but controversial — homologue of the heart with diffuse contractility[105]. These observations raise the possibility that the LCA of extant chordates had a CPF. However, contrary to urochordates and vertebrates, cephalochordates have a rather diffuse heart-like vasculature and their branchial muscles seem to develop independently of *Ebf* and *Mrf* homologues[94,106,107]. Amphioxus *Mrf* homologues seem to be expressed exclusively in somites, overlapping with the *Pax3/7* homologue[106,108], but also with the *Tbx1* homologue[102], suggesting the presence of distinct Tbx1[+], Pax3/7[+], Mrf[+] somitic and Tbx1[+], Pax3/7[−], Mrf[−] pharyngeal mesodermal domains in ancient chordates.

Branchiomeric-like muscles, such as the cephalochordate oral, velar and pterygial muscles (Fig. 4), thus probably predate the origin of a CPF as defined in urochordates and vertebrates (Fig. 3). Comparative anatomical studies suggest that the pterygial and orovelar muscles of adult amphioxus probably correspond to the atrial and oral siphon muscles of urochordates, respectively (Fig. 4). Remarkably, the ascidian oral siphon muscles (Fig. 4), which control mouth movements in post-metamorphic animals, do not derive from cardiopharyngeal progenitors[93,109,110] (Fig. 2). This is in contrast with the anterior oral muscles controlling mouth movements and in particular jaw opening (first (mandibular) arch muscles) in gnathostomes, which are CPF derivatives (Fig. 2). Comparative studies of basal chordates, including that of the fossil *Haikouella*, suggested that their pharyngeal arch series started with the second (hyoid) arch and that only during early vertebrate evolution did parts of the anterior mesoderm become incorporated into the pharyngeal series by forming a new, *Hox*-independent first arch[111,112]. Therefore, it is possible that the incorporation of the more anterior (first) arch in this series during vertebrate evolution was accompanied by integration of the associated oral

and velar muscles into the CPF. This evolutionary scenario implies that the amphioxus orovelar muscles and urochordate oral siphon muscles may be homologous to the cyclostome orovelar muscles and gnathostome mandibular muscles, which could potentially explain why these muscles are derived from the CPF only in vertebrates.

Bilaterian roots of the cardiopharyngeal network

We have argued that the presence of a CPF, with dual cardiac and skeletal myogenic capacity, is probably a synapomorphy of olfactores (a derived feature shared by urochordates and vertebrates; Figs 2, 3). This argument raises the question: do the developmental, cellular and/or molecular units that form the CPF network of olfactores have even deeper evolutionary origins? Ambulacraria (echinoderms and hemichordates) is the sister group of chordates (Fig. 3). Hemichordates possess well-defined serial gill slits and a heart–kidney complex located in the anterior-most body part (proboscis)[113]. Serially arranged pharyngeal gill openings have associated muscles in enteropneust-type hemichordates, but this musculature seems to be developmentally, anatomically and histologically distinct from the chordate branchiomeric musculature[99]. Moreover, the *Tbx1* homologue of *Saccoglossus kowalevskii*, an enteropneust hemichordate, is not expressed in the mesodermal core of the pharyngeal pouches[114], suggesting that *Tbx1* expression in pharyngeal mesoderm is a chordate synapomorphy. Further studies of ambulacrarians will test this hypothesis.

Among non-deuterostome animals, nematodes lack a heart and a defined circulatory system, but possess pharyngeal muscles that contract rhythmically, exhibit electrical activity similar to mammalian cardiomyocytes, and require *ceh-22*, the homologue of *Nkx2-5* (refs 9, 21, 22, 41, 42, 115). Flies lack anatomical structures that are comparable with the chordate pharyngeal apparatus, but the *Drosophila* homologues of *Tbx1*,

Nkx2-5, *Isl*, *Ebf* and *Mrf/MyoD* variably contribute to visceral, larval and adult skeletal and/or heart muscle specification[116–121]. The diversity of myogenic networks driving muscle identity and differentiation in flies is reminiscent of the heterogeneity of myogenic origins and programs operating in the vertebrate head. Furthermore, visceral and dorsal larval muscles in *Drosophila* develop from mesoderm in proximity to the dorsal vessel or fly heart. It is therefore conceivable that many features of the CPF gene regulatory network predate the advent of chordates and, moreover, that this regulatory circuitry preceded the emergence of the well-studied myogenic hierarchies controlling vertebrate somitic muscle development.

Evolvable cardiopharyngeal units

Here, we summarize our arguments for the origins and diversification of the CPF (Fig. 3). Filter-feeding early chordates, endowed with serial gill slits inherited from deuterostome ancestors, already had gill-associated branchiomeric, or at least branchiomeric-like, muscles (Fig. 4). A well-defined CPF then probably appeared in the olfactores. Ancestral vertebrates uncoupled myogenic specification and differentiation, thus increasing the population of cardiopharyngeal progenitors. This facilitated the emergence of cardiac chambers by progressive addition of progenitor cells to the growing heart tube during development. It also allowed for the expansion and diversification of branchiomeric muscles, contributing to increased muscularization of the pharyngeal apparatus that was essential for the transition to a predatory lifestyle. The latter was made possible by olfactores' ancestral association between branchiomeric muscles and Dlx⁺ ectoderm cells. Elaboration of this interaction permitted coevolution of the branchiomeric musculature with the newly formed neural crest-derived craniofacial skeleton, linking the novel neural-crest-derived skeletal patterns with distinct branchiomeric muscles.

We propose that the heart and atrial siphon muscle gene network seen in the urochordate *C. intestinalis* illustrates the basic ontogenetic motif underlying the specification of the vertebrate CPF[95], and suggest three ways in which this blueprint was modified to produce the vast diversity of cardiopharyngeal patterns in vertebrates: the ontogenetic motif could be deployed in multiple independent embryonic progenitors; any given progenitor could self-renew, thus being transiently amplified, before generating distinct heart, in contrast with branchiomeric, muscle precursors

and any given cell could migrate and/or be passively displaced and resume cardiopharyngeal development in different locations on receipt of appropriate signals. In contrast to their ascidian counterparts, vertebrate Tbx1⁺ and Isl1⁺ cardiopharyngeal progenitors remain in an elusive niche in which they self-renew to produce SHF-derived heart precursors. During pharyngeal morphogenesis, these emerge sequentially to produce right ventricular and outflow tract cardiomyocytes. Conceivably, multiple independent cardiopharyngeal lineages developing in series may contribute to divergent cardiac and branchiomeric myogenic cell fates along the anterior–posterior pharyngeal mesoderm of vertebrates. This hypothesis is consistent with the observation that subsets of cardiac and branchiomeric muscles are more closely related to each other than to other heart and head muscles (Fig. 1)[36,122,123]. Future experiments will determine whether anteroposterior patterning of the CPF precedes segmentation of the pharyngeal region during arch morphogenesis.

General remarks and future directions

The CPF is a new paradigm to be reckoned with, and should take centre stage along with neural crest and cranial placodes when considering the origin of the vertebrate head. Importantly, novel insights from comparative, phylogenomic and developmental genetics studies have uncovered the deep evolutionary origins of the CPF, branchiomeric muscles, placodes and neural crest cells. Like vertebrates, urochordates have a CPF that gives rise to the FHF, SHF and branchiomeric muscles; moreover, apart from their neural-crest-like cells and placodes, at least some pelagic urochordates have highly developed brains[124]. Data obtained after Gans and Northcutt's new head hypothesis thus call into question the clear distinction between vertebrates and other animals, and show that the 'new' head arose instead by elaboration and modification of existing tissues, cell populations and gene networks through evolutionary 'tinkering'. This revelation supports the proposal[125] that the conventional view of vertebrates evolving from brainless ascidian-like filter-feeders through a progressive increase in complexity and emergence of several *de novo* structures, with no evolutionary losses or reversions, is an oversimplification. These data also emphasize the heterogeneity and complex developmental and evolutionary history of vertebrate hearts and heads, blurring the interface between head and trunk, extraocular and branchiomeric, and skeletal and
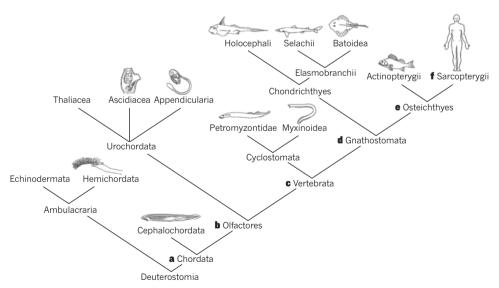
Figure 3 | Some of the synapomorphies of the Chordata and its subgroups, according to our own data and review of the literature. a, Somites and branchiomeric muscles. **b,** Placodes, neural-crest-like cells and cardiopharyngeal field (CPF) (although within invertebrates, conclusive evidence for these features was only reported in urochordates, some of these features may have been already present in the last common ancestor of extant chordates) giving rise to first- and second-heart-field-derived parts of the heart and to branchiomeric muscles (possibly not all of them, that is, inclusion of oral/velar muscles into CPF might have occurred during vertebrate

evolution). **c,** Skull, cardiac chambers, and differentiation of epibranchial and hypobranchial somitic muscles. **d,** Jaws and differentiation between hypaxial and epaxial somitic musculature; paired appendages and fin muscles; origin of the branchiomeric muscle cucullaris. **e,** Loss of epibranchial muscles; cucullaris divided into levatores arcuum branchialium (going to pharyngeal arches) and protractor pectoralis (going to pectoral girdle), an exaptation that later allowed the emergence of the tetrapod neck. **f,** Within sarcopterygians, the protractor pectoralis gave rise to the amniote neck muscles trapezius and sternocleidomastoideus.
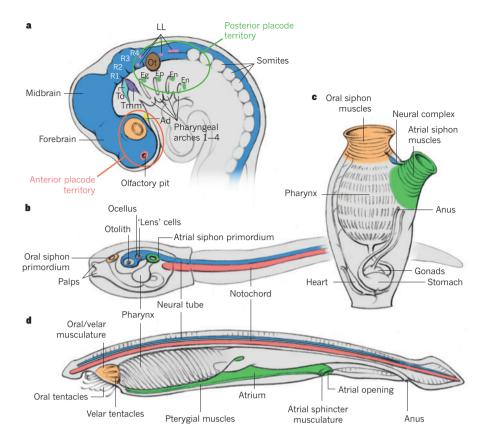
Figure 4 | Homology hypotheses of placodes and branchiomeric muscles within chordates. a, Location of ectodermal placodes in the vertebrate head according to Graham and Shimeld's[3] hypothesis (anterior to the left): olfactory placode or pit (red) at the tip of the forebrain; lens placodes (orange) form posteriorly as part of eye; adenohypophyseal placode (Ad, yellow) lies ventrally to forebrain; trigeminal placodes form alongside the anterior hindbrain at the levels of rhombomeres 1 and 2 (R1 and R2), the anterior one being the ophthalmic placode (To, light blue) and the posterior one the maxillomandibular placode (Tmm, purple); otic placode (Ot, brown) forms opposite the central domain of hindbrain; lateral line placodes (LL, pink) form anteriorly and posteriorly to otic placode; epibranchial placodes (green) — geniculate (Eg), petrosal (Ep) and nodose (En) — form as part of pharyngeal series. Forebrain, midbrain and R1–4, and neural tube are shown in dark blue. **b**, Urochordate tadpole-like larva (anterior to the left). The notochord is in red and two siphon primordia are in green and orange, with putative relationships to the anterior and posterior placode territories shown in **a**. **c**, Adult urochordate showing siphon primordia after metamorphosis. **d**, Adult cephalochordate showing the urochordate–cephalochordate muscle homology hypotheses proposed in the present Review. Figures based on images from refs 3, 22, 105.

cardiac myogenesis. Adult postcranial structures, including the heart and part of the neck musculature, include cells derived from the CPF (Fig. 1); reciprocally, cephalic structures such as the tongue and infrahyoid muscles arise from somitic primordia located in the trunk. The discovery of the CPF therefore provides a more complete, and complex, view of the origin and early evolution of the vertebrate head.

However, many questions remain. For example, how is the multipotency of branchiomeric and cardiac myocyte progenitor cells encoded in the CPF, and is there a defined molecular common niche in which these multipotent progenitor cells arise? How, and during what stages, are progenitor cell populations that give rise to different regions of the heart and head muscles specified in pharyngeal mesoderm? Recognition of the CPF also sets the stage for future discoveries in human medicine (Fig. 1). An important question is why many myopathies preferentially affect a specific subset of muscles, and whether these aetiologies are linked to the disparate embryonic histories of these muscles. As already noted, the clinical features of DiGeorge syndrome — one of the most common human congenital syndromes — include cardiovascular and craniofacial birth defects, highlighting the frequent link between these defects owing to their anatomical proximity during early embryogenesis and overlapping progenitor populations[9,21,42]. Therefore, the studies and data discussed here open promising new directions for biomedical research and the advancement of public health. For instance, future meta-analyses may reveal pathological relationships between specific branchiomeric muscles and regional congenital heart defects. The field of evolutionary developmental biology has progressed remarkably over the three decades since the new head hypothesis was published. With the recent revolutionary discoveries and more exciting work already begun, the field is poised to move ahead anew.

Note added in proof: A paper has been published while the current Review was in press reporting the identification of a third group of bilateral common heart and skeletal muscle progenitor cells within the murine CPM. Using retrospective lineage analysis, cucullaris-derived neck muscles, the trapezius and sternocleidomastoid, were shown to be clonally related to myocardium at the venous pole of the heart, derived from the posterior SHF. These findings reinforce the hypothesis of a branchiomeric origin of these neck muscles (F. Lescroart *et al*. Clonal analysis reveals a common origin between nonsomite-derived neck muscles and heart myocardium. *Proc. Natl Acad. Sci. USA* **112,** 1446–1451; 2015). ∎

1. Gans, C. & Northcutt, R. G. Neural crest and the origin of vertebrates: a new head. *Science* **220,** 268–273 (1983).
   This highly influential paper argued that the evolution of head structures derived from neural crest and cranial placodes had a crucial role in the transition to early vertebrates.
2. Patthey, C., Schlosser, G. & Shimeld, S. M. The evolutionary history of vertebrate cranial placodes — I: cell type evolution. *Dev. Biol.* **389,** 82–97 (2014).
3. Graham, A. & Shimeld, S. M. The origin and evolution of the ectodermal placodes. *J. Anat.* **222,** 32–40 (2013).
4. Northcutt, R. G. The new head hypothesis revisited. *J. Exp. Zool. B Mol. Dev. Evol.* **304B,** 274–297 (2005).
5. Kuratani, S. Evolution. A muscular perspective on vertebrate evolution. *Science* **341,** 139–140 (2013).
6. Trinajstic, K. *et al.* Fossil musculature of the most primitive jawed vertebrates. *Science* **341,** 160–164 (2013).
7. Meilhac, S. M., Esner, M., Kelly, R. G., Nicolas, J. F. & Buckingham, M. E. The clonal origin of myocardial cells in different regions of the embryonic mouse heart. *Dev. Cell* **6,** 685–698 (2004).
8. Kelly, R. G. The second heart field. *Curr. Top. Dev. Biol.* **100,** 33–65 (2012).
9. Tzahor, E. & Evans, S. M. Pharyngeal mesoderm development during embryogenesis: implications for both heart and head myogenesis. *Cardiovasc. Res.* **91,** 196–202 (2011).
10. Kelly, R. G., Brown, N. A. & Buckingham, M. E. The arterial pole of the mouse heart forms from Fgf10-expressing cells in pharyngeal mesoderm. *Dev. Cell* **1,** 435–440 (2001).
    Discovery of the mammalian SHF, demonstrating that myocardium at the arterial pole of the heart originates in adjacent pharyngeal mesoderm.
11. Mjaatvedt, C. H. *et al.* The outflow tract of the heart is recruited from a novel heart-forming field. *Dev. Biol.* **238,** 97–109 (2001).
12. Waldo, K. L. *et al.* Conotruncal myocardium arises from a secondary heart field. *Development* **128,** 3179–3188 (2001).
13. Nathan, E. *et al.* The contribution of Islet1-expressing splanchnic mesoderm cells to distinct branchiomeric muscles reveals significant heterogeneity in head muscle development. *Development* **135,** 647–657 (2008).
    This article provides a definition of the contribution of pharyngeal mesoderm to branchiomeric muscles in both chick and mouse embryos.
14. Mesbah, K. *et al.* Identification of a *Tbx1/Tbx2/Tbx3* genetic pathway governing pharyngeal and arterial pole morphogenesis. *Hum. Mol. Genet.* **21,** 1217–1229 (2012).
15. Tirosh-Finkel, L., Elhanany, H., Rinon, A. & Tzahor, E. Mesoderm progenitor cells

of common origin contribute to the head musculature and the cardiac outflow tract. *Development* **133**, 1943–1953 (2006).
This article demonstrates, using fate-mapping and experimental manipulation in the avian embryo, that cranial mesoderm gives rise both to head muscles and outflow tract myocardium.

16. Tzahor, E. & Lassar, A. B. Wnt signals from the neural tube block ectopic cardiogenesis. *Genes Dev.* **15**, 255–260 (2001).
17. Noden, D. M. & Trainor, P. A. Relations and interactions between cranial mesoderm and neural crest populations. *J. Anat.* **207**, 575–601 (2005).
18. Hutson, M. R. & Kirby, M. L. Neural crest and cardiovascular development: a 20-year perspective. *Birth Defects Res. C Embryo Today* **69**, 2–13 (2003).
19. Rinon, A. *et al.* Cranial neural crest cells regulate head muscle patterning and differentiation during vertebrate embryogenesis. *Development* **134**, 3065–3075 (2007).
20. Bothe, I. & Dietrich, S. The molecular setup of the avian head mesoderm and its implication for craniofacial myogenesis. *Dev. Dynam.* **235**, 2845–2860 (2006).
21. Grifone, R. & Kelly, R. G. Heartening news for head muscle development. *Trends Genet.* **23**, 365–369 (2007).
22. Sambasivan, R., Kuratani, S. & Tajbakhsh, S. An eye on the head: the development and evolution of craniofacial muscles. *Development* **138**, 2401–2415 (2011).
23. Cai, C. L. *et al.* Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Dev. Cell* **5**, 877–889 (2003).
24. Harel, I. *et al.* Distinct origins and genetic programs of head muscle satellite cells. *Dev. Cell* **16**, 822–832 (2009).
This article demonstrates the diversity of lineages constituting craniofacial skeletal muscles and their associated satellite cells using a series of Cre lines to genetically trace trunk and cranial myogenic progenitor cells, leading to an Isl1-lineage-based definition of CPF-derived craniofacial muscles.
25. Dodou, E., Verzi, M. P., Anderson, J. P., Xu, S. M. & Black, B. L. Mef2c is a direct transcriptional target of ISL1 and GATA factors in the anterior heart field during mouse embryonic development. *Development* **131**, 3931–3942 (2004).
26. Watanabe, Y. *et al.* Fibroblast growth factor 10 gene regulation in the second heart field by Tbx1, Nkx2–5, and Islet1 reveals a genetic switch for down-regulation in the myocardium. *Proc. Natl Acad. Sci. USA* **109**, 18273–18280 (2012).
27. Prall, O. W. *et al.* An Nkx2–5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation. *Cell* **128**, 947–959 (2007).
28. Scambler, P. J. 22q11 deletion syndrome: a role for TBX1 in pharyngeal and cardiovascular development. *Pediatr. Cardiol.* **31**, 378–390 (2010).
29. Liao, J. *et al.* Identification of downstream genetic pathways of Tbx1 in the second heart field. *Dev. Biol.* **316**, 524–537 (2008).
30. Chen, L., Fulcoli, F. G., Tang, S. & Baldini, A. Tbx1 regulates proliferation and differentiation of multipotent heart progenitors. *Circ. Res.* **105**, 842–851 (2009).
31. Hami, D., Grimes, A. C., Tsai, H. J. & Kirby, M. L. Zebrafish cardiac development requires a conserved secondary heart field. *Development* **138**, 2389–2398 (2011).
32. Kelly, R. G., Jerome-Majewska, L. A. & Papaioannou, V. E. The del22q11.2 candidate gene *Tbx1* regulates branchiomeric myogenesis. *Hum. Mol. Genet.* **13**, 2829–2840 (2004).
This paper reports the genetic identification of *Tbx1* as a regulator of craniofacial myogenesis in mice, supporting the existence of distinct upstream regulatory hierarchies controlling head and trunk myogenesis.
33. Kong, P. *et al.* Tbx1 is required autonomously for cell survival and fate in the pharyngeal core mesoderm to form the muscles of mastication. *Hum. Mol. Genet.* **23**, 4215–4231 (2014).
34. Castellanos, R., Xie, Q., Zheng, D., Cvekl, A. & Morrow, B. E. Mammalian TBX1 preferentially binds and regulates downstream targets via a tandem T-site repeat. *PLoS ONE* **9**, e95151 (2014).
35. Harel, I. *et al.* Pharyngeal mesoderm regulatory network controls cardiac and head muscle morphogenesis. *Proc. Natl Acad. Sci. USA* **109**, 18839–18844 (2012).
36. Lescroart, F. *et al.* Clonal analysis reveals common lineage relationships between head muscles and second heart field derivatives in the mouse embryo. *Development* **137**, 3269–3279 (2010).
This retrospective lineage analysis provides evidence for the existence of common progenitor cells in the mouse embryo that give rise to myocardium of the right ventricle and first-arch-derived muscles, and to the arterial pole of the heart and second-arch-derived muscles.
37. Romer, A. S. & Parson, T. S. *The Vertebrate Body* (Saunder's College Publishing, 1977).
38. Diogo, R. & Abdala, V. *Muscles of Vertebrates: Comparative Anatomy, Evolution, Homologies and Development* (CRC, 2010).
This monograph provides an overview on the comparative anatomy, evolution and homologies of the head and limb muscles in all major extant vertebrate groups with special focus on the developmental and evolutionary history of the muscles of *Homo sapiens*.
39. Devine, W. P., Wythe, J. D., George, M., Koshiba-Takeuchi, K. & Bruneau, B. G. Early patterning and specification of cardiac progenitors in gastrulating mesoderm. *eLife* **3**, e03848 (2014).
40. Lescroart, F. *et al.* Early lineage restriction in temporally distinct populations of Mesp1 progenitors during mammalian heart development. *Nature Cell Biol.* **16**, 829–840 (2014).
41. Olson, E. N. Gene regulatory networks in the evolution and development of the heart. *Science* **313**, 1922–1927 (2006).

42. Tzahor, E. Heart and craniofacial muscle development: a new developmental theme of distinct myogenic fields. *Dev. Biol.* **327**, 273–279 (2009).
43. Diogo, R. & Wood, B. A. *Comparative Anatomy and Phylogeny of Primate Muscles and Human Evolution* (CRC, 2012).
44. Wachtler, F. & Jacob, M. Origin and development of the cranial skeletal muscles. *Bibl. Anat.* **1986**, 24–46 (1986).
45. Noden, D. M. The embryonic origins of avian cephalic and cervical muscles and associated connective tissues. *Am. J. Anat.* **168**, 257–276 (1983).
46. Noden, D. M. & Francis-West, P. The differentiation and morphogenesis of craniofacial muscles. *Dev. Dynam.* **235**, 1194–1218 (2006).
47. Diogo, R., Hinits, Y. & Hughes, S. M. Development of mandibular, hyoid and hypobranchial muscles in the zebrafish: homologies and evolution of these muscles within bony fishes and tetrapods. *BMC Dev. Biol.* **8**, 24 (2008).
48. Diogo, R., Abdala, V., Lonergan, N. & Wood, B. A. From fish to modern humans — comparative anatomy, homologies and evolution of the head and neck musculature. *J. Anat.* **213**, 391–424 (2008).
49. Kuraku, S., Hoshiyama, D., Katoh, K., Suga, H. & Miyata, T. Monophyly of lampreys and hagfishes supported by nuclear DNA-coded genes. *J. Mol. Evol.* **49**, 729–735 (1999).
50. Delarbre, C., Gallut, C., Barriel, V., Janvier, P. & Gachelin, G. Complete mitochondrial DNA of the hagfish, *Eptatretus burgeri*: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly. *Mol. Phylogenet. Evol.* **22**, 184–192 (2002).
51. Delarbre, C. *et al.* The complete nucleotide sequence of the mitochondrial DNA of the agnathan *Lampetra fluviatilis*: bearings on the phylogeny of cyclostomes. *Mol. Biol. Evol.* **17**, 519–529 (2000).
52. Heimberg, A. M., Cowper-Sal-lari, R., Semon, M., Donoghue, P. C. & Peterson, K. J. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc. Natl Acad. Sci. USA* **107**, 19379–19383 (2010).
53. Ziermann, J. M., Miyashita, T. & Diogo, R. Cephalic muscles of Cyclostomes (hagfishes and lampreys) and Chondrichthyes (sharks, rays and holocephalans): comparative anatomy and early evolution of the vertebrate head. *Zool. J. Linn. Soc.* **172**, 771–802 (2014).
54. Adachi, N. & Kuratani, S. Development of head and trunk mesoderm in the dogfish, *Scyliorhinus torazame*: I. Embryology and morphology of the head cavities and related structures. *Evol. Dev.* **14**, 234–256 (2012).
55. Adachi, N., Takechi, M., Hirai, T. & Kuratani, S. Development of the head and trunk mesoderm in the dogfish, *Scyliorhinus torazame*: II. Comparison of gene expression between the head mesoderm and somites with reference to the origin of the vertebrate head. *Evol. Dev.* **14**, 257–276 (2012).
56. Kuratani, S., Adachi, N., Wada, N., Oisi, Y. & Sugahara, F. Developmental and evolutionary significance of the mandibular arch and prechordal/premandibular cranium in vertebrates: revising the heterotopy scenario of gnathostome jaw evolution. *J. Anat.* **222**, 41–55 (2013).
57. Kusakabe, R., Kuraku, S. & Kuratani, S. Expression and interaction of muscle-related genes in the lamprey imply the evolutionary scenario for vertebrate skeletal muscle, in association with the acquisition of the neck and fins. *Dev. Biol.* **350**, 217–227 (2011).
58. Kokubo, N. *et al.* Mechanisms of heart development in the Japanese lamprey, *Lethenteron japonicum*. *Evol. Dev.* **12**, 34–44 (2010).
59. Onimaru, K., Shoguchi, E., Kuratani, S. & Tanaka, M. Development and evolution of the lateral plate mesoderm: comparative analysis of amphioxus and lamprey with implications for the acquisition of paired fins. *Dev. Biol.* **359**, 124–136 (2011).
60. Sauka-Spengler, T., Le Mentec, C., Lepage, M. & Mazan, S. Embryonic expression of *Tbx1*, a DiGeorge syndrome candidate gene, in the lamprey *Lampetra fluviatilis*. *Gene Expr. Patterns* **2**, 99–103 (2002).
61. Tiecke, E. *et al.* Identification and developmental expression of two *Tbx1/10*-related genes in the agnathan *Lethenteron japonicum*. *Dev. Genes Evol.* **217**, 691–697 (2007).
62. Simões-Costa, M. S. *et al.* The evolutionary origin of cardiac chambers. *Dev. Biol.* **277**, 1–15 (2005).
63. Moorman, A. F. & Christoffels, V. M. Cardiac chamber formation: development, genes, and evolution. *Physiol. Rev.* **83**, 1223–1267 (2003).
64. Ziermann, J. M. & Diogo, R. Cranial muscle development in the model organism *Ambystoma mexicanum*: implications for tetrapod and vertebrate comparative and evolutionary morphology and notes on ontogeny and phylogeny. *Anat. Rec. (Hoboken)* **296**, 1031–1048 (2013).
65. Matsuoka, T. *et al.* Neural crest origins of the neck and shoulder. *Nature* **436**, 347–355 (2005).
66. Ziermann, J. M. & Diogo, R. Cranial muscle development in frogs with different developmental modes: direct development versus biphasic development. *J. Morphol.* **275**, 398–413 (2014).
67. Shearman, R. M. & Burke, A. C. The lateral somitic frontier in ontogeny and phylogeny. *J. Exp. Zool. B Mol. Dev. Evol.* **312**, 603–612 (2009).
68. Minchin, J. E. *et al.* Oesophageal and sternohyal muscle fibres are novel Pax3-dependent migratory somite derivatives essential for ingestion. *Development* **140**, 2972–2984 (2013).
69. Abdala, V. & Diogo, R. Comparative anatomy, homologies and evolution of the pectoral and forelimb musculature of tetrapods with special attention to extant limbed amphibians and reptiles. *J. Anat.* **217**, 536–573 (2010).
70. Edgeworth, F. H. *The Cranial Muscles of Vertebrates* (The University Press, Cambridge 1935).
This 80-year-old publication continues to be the most complete compendium on the anatomical development of the head muscles of vertebrates.

71. Piotrowski, T. & Nusslein-Volhard, C. The endoderm plays an important role in patterning the segmented pharyngeal region in zebrafish (*Danio rerio*). *Dev. Biol.* **225,** 339–356 (2000).
72. Noden, D. M. & Schneider, R. A. Neural crest cells and the community of plan for craniofacial development: historical debates and current perspectives. *Adv. Exp. Med. Biol.* **589,** 1–23 (2006).
73. Theis, S. *et al*. The occipital lateral plate mesoderm is a novel source for vertebrate neck musculature. *Development* **137,** 2961–2971 (2010).
74. Gegenbaur, C. *Elements of Comparative Anatomy* (Macmillan, 1878).
75. Gillis, J. A., Dahn, R. D. & Shubin, N. H. Shared developmental mechanisms pattern the vertebrate gill arch and paired fin skeletons. *Proc. Natl Acad. Sci. USA* **106,** 5720–5724 (2009).
76. Putnam, N. H. *et al*. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453,** 1064–1071 (2008).
77. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439,** 965–968 (2006).
78. Butler, A. B. The serial transformation hypothesis of vertebrate origins: comment on "The new head hypothesis revisited". *J. Exp. Zool. B Mol. Dev. Evol.* **306,** 419–424 (2006).
79. Gans, C. Stages in the origin of vertebrates: analysis by means of scenarios. *Biol. Rev. Camb. Philos. Soc.* **64,** 221–268 (1989).
80. Mazet, F. *et al*. Molecular evidence from *Ciona intestinalis* for the evolutionary origin of vertebrate sensory placodes. *Dev. Biol.* **282,** 494–508 (2005).
81. Mazet, F. & Shimeld, S. M. Molecular evidence from ascidians for the evolutionary origin of vertebrate cranial sensory placodes. *J. Exp. Zool. B Mol. Dev. Evol.* **304,** 340–346 (2005).
82. Wagner, E. & Levine, M. FGF signaling establishes the anterior border of the *Ciona* neural tube. *Development* **139,** 2351–2359 (2012).
83. Christiaen, L., Bourrat, F. & Joly, J. S. A modular *cis*-regulatory system controls isoform-specific pitx expression in ascidian stomodaeum. *Dev. Biol.* **277,** 557–566 (2005).
84. Christiaen, L. *et al*. *Pitx* genes in Tunicates provide new molecular insight into the evolutionary origin of pituitary. *Gene* **287,** 107–113 (2002).
85. Abitua, P. B., Wagner, E., Navarrete, I. A. & Levine, M. Identification of a rudimentary neural crest in a non-vertebrate chordate. *Nature* **492,** 104–107 (2012).
86. Satou, Y., Imai, K. S. & Satoh, N. The ascidian *Mesp* gene specifies heart precursor cells. *Development* **131,** 2533–2541 (2004).
87. Davidson, B., Shi, W. & Levine, M. Uncoupling heart cell specification and migration in the simple chordate *Ciona intestinalis*. *Development* **132,** 4811–4818 (2005).
88. Christiaen, L. *et al*. The transcription/migration interface in heart precursors of *Ciona intestinalis*. *Science* **320,** 1349–1352 (2008).
89. Davidson, B., Shi, W., Beh, J., Christiaen, L. & Levine, M. FGF signaling delineates the cardiac progenitor field in the simple chordate, *Ciona intestinalis*. *Genes Dev.* **20,** 2728–2738 (2006).
90. Beh, J., Shi, W., Levine, M., Davidson, B. & Christiaen, L. FoxF is essential for FGF-induced migration of heart progenitor cells in the ascidian *Ciona intestinalis*. *Development* **134,** 3297–3305 (2007).
91. Christiaen, L., Stolfi, A. & Levine, M. BMP signaling coordinates gene expression and cell migration during precardiac mesoderm development. *Dev. Biol.* **340,** 179–187 (2010).
92. Ragkousi, K., Beh, J., Sweeney, S., Starobinska, E. & Davidson, B. A single GATA factor plays discrete, lineage specific roles in ascidian heart development. *Dev. Biol.* **352,** 154–163 (2011).
93. Stolfi, A. *et al*. Early chordate origins of the vertebrate second heart field. *Science* **329,** 565–568 (2010).
   This article reports the discovery of the CPF in *C. intestinalis* using dynamic imaging and genetics, revealing striking genetic similarities with vertebrate pharyngeal mesoderm giving rise to head muscles and SHF-derived parts of the heart.
94. Tolkin, T. & Christiaen, L. Development and evolution of the ascidian cardiogenic mesoderm. *Curr. Top. Dev. Biol.* **100,** 107–142 (2012).
95. Wang, W., Razy-Krajka, F., Siu, E., Ketcham, A. & Christiaen, L. NK4 antagonizes Tbx1/10 to promote cardiac versus pharyngeal muscle fate in the ascidian second heart field. *PLoS Biol.* **11,** e1001725 (2013).
   This paper identified an ontogenetic motif regulating cardiac and pharyngeal skeletal muscle development in *C. intestinalis* through asymmetric cell division events and antagonistic interactions between conserved master regulators of cardiopharyngeal fate.
96. Razy-Krajka, F. *et al*. Collier/OLF/EBF-dependent transcriptional dynamics control pharyngeal muscle specification from primed cardiopharyngeal progenitors. *Dev. Cell* **29,** 263–276 (2014).
   This paper demonstrated that the multipotent cardiopharyngeal progenitors of *C. intestinalis* are multilineage primed and activate both early heart and pharyngeal muscle regulators that segregate to their corresponding precursors following asymmetric cell divisions.
97. Harafuji, N., Keys, D. N. & Levine, M. Genome-wide identification of tissue-specific enhancers in the *Ciona* tadpole. *Proc. Natl Acad. Sci. USA* **99,** 6802–6805 (2002).
98. Heude, E. *et al*. Jaw muscularization requires Dlx expression by cranial neural crest cells. *Proc. Natl Acad. Sci. USA* **107,** 11441–11446 (2010).
99. Yasui, K., Kaji, T., Morov, A. R. & Yonemura, S. Development of oral and branchial muscles in lancelet larvae of *Branchiostoma japonicum*. *J. Morphol.* **275,** 465–477 (2014).
100. Goldschmidt, R. Amphioxides. *Wiss Ergeb Dtsch Tiefsee-Expedition* [in German] **12,** 1–92 (1905).
101. Holland, N. D., Venkatesh, T. V., Holland, L. Z., Jacobs, D. K. & Bodmer, R. *AmphiNk2-tin*, an amphioxus homeobox gene expressed in myocardial progenitors: insights into evolution of the vertebrate heart. *Dev. Biol.* **255,** 128–137 (2003).
102. Mahadevan, N. R., Horton, A. C. & Gibson-Brown, J. J. Developmental expression of the amphioxus *Tbx1/10* gene illuminates the evolution of vertebrate branchial arches and sclerotome. *Dev. Genes Evol.* **214,** 559–566 (2004).
103. Jackman, W. R., Langeland, J. A. & Kimmel, C. B. *islet* reveals segmentation in the Amphioxus hindbrain homolog. *Dev. Biol.* **220,** 16–26 (2000).
104. Belgacem, M. R., Escande, M. L., Escriva, H. & Bertrand, S. Amphioxus Tbx6/16 and Tbx20 embryonic expression patterns reveal ancestral functions in chordates. *Gene Expr. Patterns* **11,** 239–243 (2011).
105. Willey, A. *Amphioxus and the Ancestery of the Vertebrates* (Macmillan, 1894).
106. Schubert, M., Meulemans, D., Bronner-Fraser, M., Holland, L. Z. & Holland, N. D. Differential mesodermal expression of two amphioxus MyoD family members (AmphiMRF1 and AmphiMRF2). *Gene Expr. Patterns* **3,** 199–202 (2003).
107. Mazet, F., Masood, S., Luke, G. N., Holland, N. D. & Shimeld, S. M. Expression of *AmphiCoe*, an amphioxus COE/EBF gene, in the developing central nervous system and epidermal sensory neurons. *Genesis* **38,** 58–65 (2004).
108. Holland, L. Z., Schubert, M., Kozmik, Z. & Holland, N. D. *AmphiPax3/7*, an amphioxus paired box gene: insights into chordate myogenesis, neurogenesis, and the possible evolutionary precursor of definitive vertebrate neural crest. *Evol. Dev.* **1,** 153–165 (1999).
109. Hirano, T. & Nishida, H. Developmental fates of larval tissues after metamorphosis in ascidian *Halocynthia roretzi*. I. Origin of mesodermal tissues of the juvenile. *Dev. Biol.* **192,** 199–210 (1997).
110. Tokuoka, M., Satoh, N. & Satou, Y. A bHLH transcription factor gene, *Twist-like1*, is essential for the formation of mesodermal tissues of *Ciona* juveniles. *Dev. Biol.* **288,** 387–396 (2005).
111. Kuratani, S. Evolution of the vertebrate jaw from developmental perspectives. *Evol. Dev.* **14,** 76–92 (2012).
112. Mallatt, J. The origin of the vertebrate jaw: neoclassical ideas versus newer, development-based ideas. *Zoolog. Sci.* **25,** 990–998 (2008).
113. Valentine, J. W. *On the Origin of Phyla* (Univ. Chicago Press, 2004).
114. Gillis, J. A., Fritzenwanker, J. H. & Lowe, C. J. A stem-deuterostome origin of the vertebrate pharyngeal transcriptional network. *Proc R. Soc. B* **279,** 237–246 (2012).
115. Haun, C., Alexander, J., Stainier, D.Y. & Okkema, P. G. Rescue of *Caenorhabditis elegans* pharyngeal development by a vertebrate heart specification gene. *Proc. Natl Acad. Sci. USA.* **95,** 5072–5075 (1998).
116. Boukhatmi, H. *et al*. An Org-1-Tup transcriptional cascade reveals different types of alary muscles connecting internal organs in *Drosophila*. *Development* **141,** 3761–3771 (2014).
117. Crozatier, M. & Vincent, A. Requirement for the *Drosophila* COE transcription factor Collier in formation of an embryonic muscle: transcriptional response to notch signalling. *Development* **126,** 1495–1504 (1999).
118. Enriquez, J., de Taffin, M., Crozatier, M., Vincent, A. & Dubois, L. Combinatorial coding of *Drosophila* muscle shape by Collier and Nautilus. *Dev. Biol.* **363,** 27–39 (2012).
119. Mann, T., Bodmer, R. & Pandur, P. The *Drosophila* homolog of vertebrate *Islet1* is a key component in early cardiogenesis. *Development* **136,** 317–326 (2009).
120. Schaub, C. & Frasch, M. Org-1 is required for the diversification of circular visceral muscle founder cells and normal midgut morphogenesis. *Dev. Biol.* **376,** 245–259 (2013).
121. Schaub, C., Nagaso, H., Jin, H. & Frasch, M. Org-1, the *Drosophila* ortholog of Tbx1, is a direct activator of known identity genes during muscle specification. *Development* **139,** 1001–1012 (2012).
122. Lescroart, F. & Meilhac, S. M. Cell lineages, growth and repair of the mouse heart. *Results Probl. Cell Differ.* **55,** 263–289 (2012).
123. Lescroart, F., Mohun, T., Meilhac, S. M., Bennett, M. & Buckingham, M. Lineage tree for the venous pole of the heart: clonal analysis clarifies controversial genealogy based on genetic tracing. *Circ. Res.* **111,** 1313–1322 (2012).
124. Lacalli, T. C. & Holland, L. Z. The developing dorsal ganglion of the salp *Thalia democratica*, and the nature of the ancestral chordate brain. *Phil. Trans. R. Soc. Lond. B* **353,** 1943–1967 (1998).
125. Gee, H. in *Major Events in Early Vertebrate Evolution: Palaeontology, Phylogeny, Genetics and Development* (ed. Ahlberg, P. E.) 1–14 (Taylor & Francis, 2001).

Evolution of vertebrates as viewed from the crest

Stephen A. Green[1], Marcos Simoes–Costa[1] & Marianne E. Bronner[1]

The origin of vertebrates was accompanied by the advent of a novel cell type: the neural crest. Emerging from the central nervous system, these cells migrate to diverse locations and differentiate into numerous derivatives. By coupling morphological and gene regulatory information from vertebrates and other chordates, we describe how addition of the neural–crest–specification program may have enabled cells at the neural plate border to acquire multipotency and migratory ability. Analysis of the topology of the neural crest gene regulatory network can serve as a useful template for understanding vertebrate evolution, including elaboration of neural crest derivatives.

The vertebrate body plan emerged in concert with extensive changes to anterior chordate morphology, including assembly of a craniofacial skeleton, expansion of the anterior neuroepithelium into a brain, reorganization of the pharynx and appearance of novel sensory systems[1–3]. The genesis of this vertebrate 'new head'[1] has been fundamentally linked to the emergence of two cell types, neural crest cells and ectodermal placodal cells. The neural crest is a transient vertebrate cell type, characterized by its site of origin within the central nervous system (CNS), multipotency, and its ability to migrate and differentiate into numerous derivatives, as diverse as cartilage, bone, melanocytes, peripheral neurons and glia[4]. Together with ectodermal placodes that give rise to the sense organs of the head (see refs 5, 6 for discussion of placode evolution), neural crest cells have contributed to the remarkable array of novel anatomies that make vertebrates unique.

Neural crest cells are unlike any other cell type, and the advent of this progenitor cell population affected chordate evolution in an unprecedented manner. Although cells with subsets of neural crest characteristics are present in invertebrate chordates, only vertebrates have a bona fide neural crest that gives rise to structural elements of the head, glia, pigment cells and neurons. Imbued with broad developmental potential and extensive migratory ability, neural crest cells have gained developmental roles at nearly all axial levels and extensively interact with many other tissues. For these reasons, the neural crest is often referred to as the fourth germ layer[7], associated with the emergence and elaboration of the vertebrate body plan[1,8,9].

In this Review, we examine the morphological and genetic features that distinguish vertebrates from other chordates, focusing on cells and tissues derived from the neural crest. We place special emphasis on contributions that resulted in the assembly of the vertebrate head, which has played a crucial part in establishment and diversification of vertebrates. We discuss the gene regulatory network (GRN) underlying the formation of the early neural crest cells that are common to all vertebrates. We then use this network, together with morphological criteria, to discuss how neural crest cells may have emerged from the putative homologues that are present in invertebrate chordates, highlighting how addition of the neural-crest-specification program may have enabled cells at the CNS border to acquire multipotency and migratory ability. In this context, we examine how studies of neural crest GRNs may clarify patterns of morphological evolution within vertebrates, including expansion of neural crest derivatives during diversification of vertebrate taxa.

Taken together, the data paint a picture of the neural crest as a malleable population that has continued to imbue the vertebrate body with novel features.

Neural–crest–related innovations in early vertebrates

Emergence of the vertebrate lineage was accompanied by acquisition of the neural crest and its novel derivatives. All vertebrates have neural crest cells that arise from the dorsal portion of the CNS, exhibit multipotency by contributing to diverse derivatives, undergo an epithelial-to-mesenchymal transition (EMT), and have extensive migratory ability. 'Premigratory' neural crest cells initially reside in or adjacent to the dorsal neural tube, the newly formed CNS, of all vertebrates[10]. These cells undergo EMT to exit the CNS and migrate to numerous sites throughout the body, where they eventually contribute to their characteristic derivatives[4] (Fig. 1a). Cell-lineage analyses have shown that many individual neural crest precursors can contribute to multiple cell types in vivo[11–13] and in vitro[14,15], and are thus 'multipotent' stem or progenitor cells.

Comparisons between the two major groups of living vertebrates, the jawed vertebrates (gnathostomes) and their sister group the cyclostomes (agnathans)[16], identify many shared, derived traits likely to have been present in the neural crest of early vertebrates[17–20]. These include pigment cells, cellular pharyngeal cartilage and specialized pharyngeal musculature, an enteric nervous system, chromaffin cells, and perhaps cardiac valves[17,21]. Recent work has identified a new neural crest derivative, pillar cells[22], that support vertebrate gill epithelia (Box 1). Because neural crest cells interact with many other tissues, they have a broad impact by modifying neuroepithelial patterning, craniofacial patterning, and cranial musculoskeletal development (Box 2).

Many early vertebrate innovations are unique to jawed vertebrates and absent in cyclostomes. Some of these traits are likely to have arisen in stem gnathostomes, the early fishes that are ancestral to the jawed vertebrates. One of these innovations is the appearance of jaws, through modification of anterior pharyngeal arches. Other major gnathostome innovations include odontoblasts that produce dentine (Box 1), paravertebral sympathetic chain ganglia[23] (Box 3) and exoskeletal armour. Although exoskeletal armour might have arisen from neural crest at cranial levels, it is likely that trunk armour instead arose from mesoderm (Box 4).

One central question in the early evolution of neural crest is the extent to which neural crest cell types are evolutionary novelties, rather than cell types (and regulatory programs) co-opted from other tissues. There are clearly some novel neural-crest-derived cell types, including

a Vertebrate neural crest development

b Vertebrate neural crest GRN

c Tunicate NC-like cell circuit
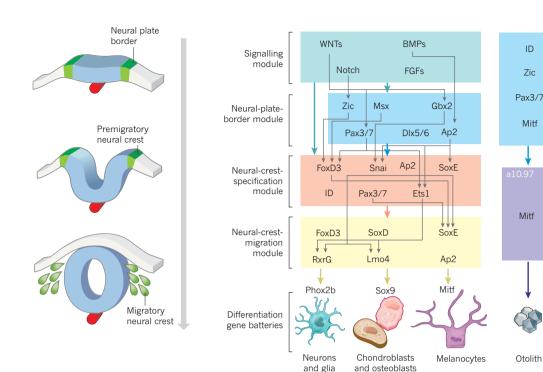
Figure 1 | Gene regulatory interactions controlling vertebrate neural crest formation and the tunicate a9.49 cell lineage. a, Different stages in neural crest formation. Neural crest cells are defined by their origin at the neural plate border, epithelial to mesenchymal transition, migratory capacity and multipotency. **b**, A neural crest gene regulatory network (GRN) endows this cell population with its unique features. This GRN is composed of different modules arranged hierarchically, which control each step of neural crest development[38]. The neural-crest-specification module seems to be missing from the neural plate border of invertebrate chordates. **c**, Regulatory circuit of a tunicate neural-crest (NC)-like pigmented cell precursor. Adapted from refs 38, 39 and based on the results from ref. 49.

pillar cells and odontoblasts, but many neural crest cell types are similar to cells in related chordates[24,25]. These cell types might either be homologous, representing a cell lineage that was co-opted and incorporated into the neural crest, or they might have arisen by convergent evolution. One example of co-option is the origin of pharyngeal cellular cartilage, probably accomplished by reuse of a program governing cellular cartilage formation in the oral region of invertebrate chordates[26]. Assessment of co-option or novelty can be aided by evaluation of GRNs that govern their formation.

A neural crest GRN is conserved across vertebrates

From a gene regulatory perspective, the body plan of all metazoans is encoded in the genome. During embryonic development, this code emerges as a complex GRN formed by transcription factors and *cis*-regulatory elements that co-operate with non-coding RNAs and epigenetic factors to pattern the body and drive development of individual elements and cell types[27]. According to this framework, the body-plan modifications observed during evolution are a direct consequence of changes in the developmental regulatory program[28].

Neural crest cells are characterized by site of origin, migratory behavior and multipotency. Importantly, they also share a molecular signature, expressing a suite of transcription factors, including *tfAP2* (ref. 29), *Snai1/2* (ref. 30), *FoxD3* (refs 31–33) and *SoxE* (refs 34, 35) genes. In particular, *FoxD3* and *SoxE* are characteristic of premigratory and early migratory neural crest cells and *SoxE* genes are crucial upstream regulators of all neural crest lineages. These transcription factors are part of the regulatory machinery that controls transcription of numerous effector genes, which together endow the neural crest with its unique properties. Interactions between transcription factors and their targets generate a GRN that controls neural crest formation, from induction at the neural plate border to differentiation into distinct cell types[36–39] (Fig. 1b).

The architecture of the neural crest GRN is thought to underlie the features observed in this cell population, such as multipotency and migratory capability. Functional experiments suggest that the neural crest GRN is comprised of distinct hierarchical levels[36,38]. First, signalling events (GRN signalling module) initiate the specification process, by inducing co-expression of transcription factors that comprise the 'neural-plate-border module'. This in turn leads to specification of bona fide neural crest cells (neural-crest-specification module), their migration from the CNS to diverse sites (neural-crest-migration module), and finally to diversification into different derivatives through the deployment of distinct differentiation gene batteries[36–39] (Fig. 1b). Each level of the neural crest GRN corresponds to a regulatory state that not only defines cell identity and behaviour at a given time point, but also drives transition to the next module of the network[40]. From an evolutionary perspective, assessing conservation of different levels of the neural crest GRN helps to identify the origin of each subcircuit and reconstruct the evolutionary history of neural crest cells[27,28]. As a result, the neural crest GRN provides a useful platform for understanding the molecular underpinnings of vertebrate evolution and how these cells may have participated in modifying vertebrate embryonic development. Neural-crest-GRN studies have indeed provided important clues regarding the establishment of the vertebrate lineage and its diversification[40–42].

Extensive work in amniotes, frogs, teleosts and cyclostomes has revealed remarkable similarities in the overall structure of the neural crest GRN, demonstrating that it is virtually the same from amniotes to cyclostomes (Fig. 1b)[8,10,19,43]. Some important species-specific differences exist, but they are likely to reflect the continuous restructuring of the GRN in individual clades. Nevertheless, expression patterns and epistatic interactions between FoxD3, SoxE, Snai1/2 and Pax3/7 transcription factors point to a very conserved module of neural crest specification[38]. The overall conservation of the neural crest GRN correlates with conservation

BOX 1

Neural crest derivatives and the vertebrate pharynx

Changes in pharyngeal patterning are central to the evolution and diversification of vertebrate groups[1,98]. Vertebrate pharyngeal arches have a similar general structure, characterized as a bilaterally symmetric series of endodermal evaginations that, with ectoderm, enclose a region of neural crest cells surrounding paraxial mesoderm[99,100]. Neural crest cells and paraxial mesoderm give rise to pharyngeal skeletal elements and musculature, respectively.

Some aspects of vertebrate pharyngeal patterning are integrated within or modified from features common to many deuterostomes. Pharyngeal segmentation is a trait of ancestral deuterostomes[101], and unambiguous pharyngeal arch homologues with similar genetic controls are present in hemichordates, cephalochordates and adult urochordates[99,101], despite being secondarily lost in echinoderms[99,102]. Pharyngeal mesoderm also has a broad phylogenetic distribution, being present throughout chordates[103,104]. Neural-crest-derived cellular cartilage of vertebrates, rather than being a novelty of vertebrates[21], instead seems to have been co-opted from cellular cartilage homologous to that present within the oral cirri of cephalochordates[26].

Although some vertebrate pharyngeal patterning stems from ancestral conditions, many novel elements arise from vertebrate neural crest cells. Modification of early neural crest development was important for generating the diversity of pharyngeal structures observed throughout vertebrates. For example, in vertebrate gills, epithelial surfaces are supported by novel neural-crest-derived cells, pillar cells, which are ancestrally shared throughout vertebrates[22]. In addition, in the transition from agnathans to gnathostomes, modifications to the anterior-most pharyngeal arch cartilages and neural-crest-modified musculature resulted in the formation of the jaws, as well as the formation of neck muscles[18,105–107].

Another vertebrate novelty associated with the pharynx and its integuments are odontodes: dental elements composed of mineral material and associated cells. In living jawed vertebrates, their formation is mediated by conserved gene regulatory subcircuits, identified by coexpression of transcription factors, including runx2 and eda/edar, among others[108], and require the inductive influence of neural-crest-derived mesenchyme. Fossil evidence suggests that odontodes emerged during the evolution of stem gnathostomes, in external dermal armour[108–110], consistent with the 'outside–in' model, which suggests that odontodes emerged first as structural elements associated with external integument, and were later incorporated into the oral cavity and pharynx. Mineralized dental elements found in conodont fossils are considered non-homologous to gnathostome teeth[109]. Both groups of living cyclostomes, lampreys and hagfish, have keratinized dental elements, but these are morphologically distinct from gnathostome teeth and are probably not homologous. Continued analysis of cyclostome dental elements might clarify whether neural crest cells played a part in their ontogeny.

---

of morphology, migratory behaviour and differentiation into multiple derivatives, establishing the neural crest as an ancient vertebrate cell type. Superimposed on the conserved basic structure of the neural crest GRN is adaptability and flexibility. During the course of evolution, differentiation modules that encode for novel derivatives, such as jaws and sympathetic ganglia, have been added to the neural crest repertoire and thus must have been added as 'plug-ins' to the GRN.

Although the core elements are highly conserved, adaptations, additions and potentially losses have occurred between species. Indeed, it is clear that the specification module of the neural crest GRN is strongly conserved within vertebrates, but there are important gene regulatory differences between jawless and jawed vertebrates that might provide interesting hints regarding the molecular roots of vertebrate morphological diversification. Extensive analysis of the lamprey neural crest GRN has revealed the notable absence of transcription factors *Ets-1* and *Twist* in the premigratory neural crest[10]. This is intriguing since Ets-1 has been shown to be essential for cranial neural crest specification in gnathostomes[34]. Instead, in the lamprey, it is expressed much later in the neural-crest-derived portion of the branchial arches and dorsal root ganglia. One possibility is that Ets-1 was added to the gnathostome neural crest specification, representing an example of a transcription factor that was co-opted from a distal level of the network to a more proximal level. However, it is also possible that it may have been selectively lost in the lamprey neural crest. Examining expression of Ets-1 in other cyclostomes and functional experiments in lampreys may help to clarify this point. Other GRN components that have crucial functions in teleosts and amphibians may have been lost or replaced in amniotes. For example, although Snai1/2 and Twist seem to be crucial for neural crest formation in frogs[44,45], they are dispensable in mice[46], perhaps due to redundant functions with other EMT factors such as Sip1 (ref. 47).

Taken together, these studies reveal that the topology of the neural crest GRN, with cells progressing through successive regulatory states from induction to differentiation, forms a useful template for understanding vertebrate evolution[36]. This GRN can also be useful for assessing the likelihood that similar cell types in other animals might be homologous to the neural crest.

Do invertebrate chordates have neural crest cells?

Deciphering how the neural crest arose as a cell type is crucial for furthering our understanding of vertebrate evolution. Tackling this problem requires deeper knowledge of deuterostome embryonic development in multiple species, with particular attention to neural-crest-like cell types in other chordates. Recent studies have described intriguing embryonic cell populations in ascidians that have some, but not all, neural crest characteristics. For example, the trunk lateral cells in the colonial tunicate *Ecteinascidia turbinata* are derived from the A7.6 lineage, which originates in the vicinity of the neural tube, undergoes migration and gives rise to pigmented cell types[48]. Similarly, in *Ciona intestinalis*, results show that the a9.49 cell lineage originates from the neural plate border and gives rise to the pigmented sensory cells of the otolith and the ocellus[49]. These cells normally translocate only a few cell diameters, whereas misexpression of Twist in this lineage results in acquisition of mesenchymal morphology and long-range migration[49]. In cephalochordates, there have been many proposed homologues of neural crest (see ref. 50 for a discussion), including a bipotential neuroepithelial precursor to pigment cells of the ocellus[50]. Further assessment of this homology will require additional analyses of amphioxus ocellus development. Cephalochordates also have an ependymal cell in the neural tube that expresses *Snail*, a homologue of *Snai1* and a neural-crest-specifier gene in vertebrates, but this cell seems to be non-migratory[51,52].

The neural crest GRN is particularly useful for understanding assessment of GRN conservation outside of vertebrates. The available molecular data obtained from embryonic cell types in tunicates and cephalochordates suggest that gene regulatory interactions that specify the neural plate border (neural-plate-border module) are deeply conserved throughout chordates[24,51] (Fig. 1c), and data from annelids suggest that this genetic program might be shared with protostomes, originating in stem bilaterians[53,54]. Similarly, the terminal differentiation programs (differentiation gene batteries) that drive the neural crest to assume definitive fates are

conserved, as exemplified by control of pigment-cell differentiation. This is expected because most of the differentiation batteries are thought to be ancient subcircuits that were co-opted by different cell types[27]. Although they are integral parts of the neural crest GRN, these neural-plate-border and differentiation subcircuits do not fully define neural crest identity in vertebrates. Proximally in the program, the neural plate border contains other cell types (neural tube and placode) in addition to neural crest, and is important for the delimitation of the neural plate. Distally, other deuterostomes have some differentiated cell types that in vertebrates can arise from neural crest: melanocytes, ectomesenchyme, autonomic neurons and glia. It has been proposed that during early vertebrate evolution, the neural-crest-specification module may have been assembled within the neural-plate-border cell lineage, interposed between the neural plate border and the distal differentiation modules of the network, to endow these cells with a full 'neural crest' phenotype.

Importantly, neural crest identity in all vertebrates is intrinsically linked to the neural-crest-specification kernel of the GRN, which endows these cells with its defining features such as multipotency, the ability to undergo EMT and migratory capacity[40]. Important genes in the specification subcircuit include *SoxE*, *FoxD* and *Snai1/2*, homologues of which are present in the genomes of invertebrate chordates[51,55]. For example, the amphioxus genome has all the transcription factors identified in the neural-crest-specifier module of the vertebrate neural crest GRN. However, only *AmphiSnail* is expressed in the putative neural crest domain[56]. Therefore, a key question is whether the neural-crest-like cells from tunicates possess this particular subcircuit. Molecular analyses suggest that tunicates and amphioxus have the neural-plate-border subcircuit[24], and thus invertebrate neural-crest-like cells may be homologous to neural-plate-border cells of vertebrates. However, although some neural-plate-specifier genes are expressed in these cells (for example, *FoxD*[49]) other crucial transcription-factor genes, notably *SoxE* genes, seem to be absent. In ascidians, it is not yet clear whether epistatic interactions between the transcription factors expressed in putative neural crest cells are similar to those observed in the vertebrate neural crest GRN (Fig. 1c). This, together with the fact that cells of the a9.49 lineage have not yet been shown to be multipotent, or to have extensive migratory capabilities, makes it more difficult to determine whether they are true neural crest homologues. Further gene-regulatory studies will be necessary to establish the relationship between these cells and the vertebrate neural crest.

As a cautionary note, there is inherent danger in assigning evolutionary relationships among cell types on the basis of molecular similarity alone, because transcription factors are reused throughout development, and are neither lineage- nor cell-type-specific. For instance, many bona fide neural crest transcription factors are expressed at the neural plate border, in later differentiation programs and in other lineages. Thus, one cannot attribute homology or lineage relationships on the basis of a few molecular markers. A more inclusive argument that includes morphological and behavioural information, expression data and, ideally, *cis*-regulatory studies[57] perhaps provides the most reliable means to establish conservation of developmental mechanisms and ascribe homology between cell populations.

Gene regulatory changes behind neural crest emergence

Radical changes of body plan, such as those that took place in early vertebrate evolution, require substantial rearrangements in the structure of developmental GRNs[27]. The emergence of the neural crest was dependent on the assembly of a specification subcircuit that allowed this cell population not only to exhibit its stereotypical behaviour, but also to drive multiple differentiation programs, resulting in its multipotent state. Understanding how a novel, complex specification subcircuit emerged during chordate evolution is a daunting task. However, observation of the neural crest GRN can provide important clues about vertebrate evolution and suggest likely scenarios for the creation of a novel cell type.

Given the deep conservation of the neural-plate-border-specification program[24], it seems reasonable to assume that this circuit was crucial for assembly of the vertebrate neural crest GRN. Because all of the neural-crest-specifier genes are present in the genomes of invertebrate chordates[58,59], it is likely that they were added to the GRN by deployment or co-option of transcription factors that were originally part of other developmental GRNs, such as the neural-plate-border subcircuit, mesodermal programs and terminal differentiation modules. According to this view, changes in their *cis*-regulatory apparatus placed the neural-crest-specifier genes downstream of the neural-plate-border program and signalling systems. Such *cis*-regulatory changes might have facilitated redeployment of neural-plate-border (*Pax3/7* and *TFAP2*) and stem-cell genes (*FoxD3*) in the specification module. For example, an amphioxus *FoxD* enhancer that recapitulates endogenous amphioxus *FoxD* expression

---

BOX 2

Role of the neural crest in signalling

Brain and facial patterning. Increased complexity in vertebrate neuroanatomy might partly stem from interactions between neural crest cells and other cell types. An example of the important role of the neural crest in expansion of the head comes from recent experiments in amniotes[111]. Surgical removal of the neural crest at forebrain to rostral hindbrain levels results in the absence of facial and skull cartilages and bones, as well as severe brain defects including anencephaly[112]. These defects can be rescued by grafting small populations of premigratory neural crest from the same axial level, but not from more caudal regions with Hox gene expression. At a molecular level, this results from production of BMP inhibitors, Gremlin and Noggin, by the rostral neural crest that in turn lead to regulation of expression of FGF8 in the anterior neural ridge (ANR). Consistent with this, implantation of FGF8 beads after neural crest ablation rescues this phenotype to restore subsequent downstream signalling events and proper head development[100,113]. FGF signalling associated with an ANR-like signalling centre is potentially present throughout deuterostomes[114,115], suggesting that neural crest cells have adopted or co-opted roles in the regulation of neural or craniofacial patterning, at least in amniotes. Examination of additional vertebrate groups might clarify when this might have arisen.

Cranial muscles and the neural crest. The vertebrate head includes muscles that control the movement of the eyes (extraocular muscles), face, jaws, throat, larynx and tongue, collectively called branchiomeric muscles[116]. Derived from unsegmented paraxial mesoderm anterior to the otic vesicle, they form under the control of a Pitx2c and Tcf21/MyoR regulatory subcircuit that seems to be conserved at least throughout the bony fishes[117,118] (Fig. 2). The neural crest is crucial for multiple stages of cranial mesoderm development, including defining the location, orientation, patterning and differentiation state of muscle precursor cells[57,106,107,116]. Mesoderm cells follow migrating neural crest cells into the pharyngeal arches[86,116]. Branchiomeric muscles initially remain in a precursor state, repressed by signals emanating from the nearby neural tube and ectoderm. Neural crest cells secrete signals that derepress myogenesis, allowing the formation of cranial myofibres[119]. These distinct myogenic regulatory sub-networks are thought to have arisen in early vertebrates concurrent with other cephalic modifications[117,119], but have also been compared with muscle precursors in the amphioxus atrium[104] and potentially with visceral musculature of protostomes[120]. Vertebrate cranial muscle patterning, differentiation and organization might require regulatory control that arose from novel interactions with the neural crest (Fig. 2).

BOX 3

Peripheral nervous system

A peripheral nervous system, including the sympathetic chain ganglia, is a common feature of all jawed vertebrates. Sympathetic ganglion cells are responsible for regulating homeostatic functions of peripheral organs. They arise from neural crest cells that migrate ventrally from the trunk neural tube to positions adjacent to the dorsal aorta, and form under the control of a gene regulatory circuit including *Phox2*, *Hand2* and *Ascl1*. These genes collaborate to promote the construction of a sympathetic neural phenotype, including production of noradrenaline. In bony fishes and tetrapods, sympathetic ganglia are connected along the anteroposterior axis through chains, but in extant chondrichyans (sharks, rays and skates) ganglia are largely separate. Cyclostomes do not seem to have a comparably organized sympathetic system, but very rare ganglion-like cells of unknown function have been identified[121]. In general, autonomic function in cyclostomes seems to be controlled directly by spinal neurons of the central nervous system[121], which is similar to the peripheral organization of amphioxus, and thus is likely to represent a primitive condition for chordates. Taken together, these data suggest that sympathetic ganglia probably evolved in stem gnathostomes, and were further elaborated in stem osteichthyes.

in somites and notochord[60] was able to drive similar expression when electroporated into chick embryos[51]. However, this enhancer failed to drive expression in the neural crest, suggesting that the novel neural crest expression domains rely on distinct gene regulatory processes that are absent in amphioxus[51]. Similarly, co-option of EMT driver genes such as *Snai2* (ref. 30) and *Sip1* (ref. 47) may have allowed the neural crest to leave the neural plate border domain. This was probably accompanied by co-option of mesenchymal gene circuits that allowed these cells to exhibit migratory behaviour.

A key feature of the neural crest is its ability to form numerous derivatives (multipotency). Mechanistically, this implies that neural crest cells are capable of deploying a variety of differentiation gene batteries depending on signalling interactions during migration and once at their final sites. Neural-crest-specifier genes from the SoxE family play a crucial part in activating differentiation programs that lead to multiple derivatives, as diverse as neurons, Schwann cells, pigment cells and cartilage[38]. Thus, a likely scenario was that a variety of differentiation gene batteries were placed downstream of the neural-crest-specification module by gain of function *cis*-regulatory changes, which placed differentiation driver genes (for example, *Mitf*, *Ascl1* or *Phox2b*) under the control of neural-crest-specifier genes. Again, examples of redeployment of such ancient differentiation gene batteries by different cell types have been described in different contexts, and are thought to be a common feature in GRN evolution[27,61]. Indeed, a recent study[26] suggests that *cis*-regulatory changes in ancestral pro-chondrocytic genes allowed for their activation in the neural crest by factors such as SoxE and Tfap2, allowing for the establishment of the vertebrate head skeleton. Thus, it is possible that the emergence of the neural-crest-specifier module served as a platform for the redeployment of multiple, pre-existing genetic subcircuits that endowed the neural crest with its defining features.

Although *cis*-regulatory changes were probably the most important events in the emergence of the neural-crest-specification module, it is also likely that changes in protein sequence had an important role therein. Neural crest cells employ a large repertoire of adhesion molecules, receptors and signalling molecules, and gene diversification and neofunctionalization might have enabled acquisition of the complex cell behaviours exhibited by the neural crest. Furthermore, recent data suggest that neofunctionalization of neural-crest-specifier genes such as *FoxD3* was

important for the emergence of this cell type[62], perhaps by mediating new protein–protein interactions and allowing for the assembly of novel, vertebrate-specific transcriptional complexes.

A role for gene duplications in early neural crest evolution

The extensive changes in gene regulation required for the evolution of the neural crest as a cell type might have been facilitated by large-scale genome duplications that took place early in the vertebrate lineage. It has long been suspected that rare, large-scale genomic rearrangements and genome-wide duplications in stem vertebrates had a key role in elaborating the vertebrate body plan[54,63–65] and increasing vertebrate complexity[66,67]. The presence of multiple homologous Hox clusters and conserved syntenic paralogy regions among jawed vertebrate chromosomes are usually taken to support the contention that there were two rounds of genome duplication during early vertebrate evolution[66]. Recent analysis of the genome of the sea lamprey (*Petromyzon marinus*) suggested that ancestors of the lamprey (and hagfish) diverged from vertebrates after these two rounds of duplication[68–70], but this is still controversial, and an alternative model suggests that there was only a single round of duplication in stem vertebrates, followed by lineage-specific segmental duplications in jawed vertebrates and cyclostomes[71]. Regardless of the precise number and timing of genome duplications, vertebrates have certainly undergone additional gene duplications relative to invertebrates, and these increases in gene number may have facilitated the evolution of vertebrate regulatory and anatomic complexity[63], potentially affecting the formation of the many novel cell types in vertebrates.

A full assessment of the extent to which gene and genome duplications have affected early vertebrate evolution remains incomplete, and is somewhat controversial[72]. One way to approach this question is to determine whether the timing of the acquisition of particular traits compares with the inferred timing of gene duplications. Many traits were thought to arise in the vertebrate stem: these include key innovations such as the addition of neural-crest-derived pharyngeal cartilages, modification of cranial muscles, the development of segmented and Hox-patterned hindbrain[57], and perhaps the beginnings of peripheral nervous organization (Fig. 2). These distinct vertebrate characters are rooted in invertebrate chordates, but seem to have been fundamentally transformed by the innovation of neural crest cells and their interactions with other cell types. Thus, the timing of the acquisition of these traits correlates nicely with inferred instances of genome duplication, although one cannot distinguish cause from effect.

Ultimately, the fundamental question is how genomic duplications affected the organization of developmental GRNs. As has been discussed[54], such duplications may cause important shifts in gene regulatory mechanisms during vertebrate evolution. Indeed, it is possible that large-scale genome duplications may have facilitated extensive changes in the *cis*-regulatory apparatus controlling the transcription of neural crest genes[73], leading to their co-option and assembly into the neural-crest-specification module. Such events might have enabled the deployment of genes, such as those that encode SoxE transcription factors, in the neural-crest-specification module. Depending on the species, *Sox8*, *Sox9* and *Sox10* have early and sometimes overlapping functions in neural crest specification, with different paralogues deployed at different times depending on the species. However, expressing at least one of the *SoxE* paralogues seems crucial for the maintenance of neural crest identity. Interestingly, it has recently been shown that *Sox10* alone is sufficient to reprogram fibroblast cells to a neural crest fate, highlighting the importance of *SoxE* genes in neural crest specification[74]. Furthermore, acquisition of migratory ability by the neural crest may have been fostered by diversification of receptors and ligands that enabled chemotactic behaviour. Genome-wide analysis shows that vertebrates have a much more complex arsenal of such molecules than do invertebrate chordates[58,75]. Thus, although the role of whole-genome duplications in neural crest evolution is still not fully understood, it

is likely that these duplications provided the neural crest with the molecular toolkit necessary for its complex behaviour.

Evolution of crest populations along the rostrocaudal axis

Neural crest cells arising from different levels of the neural axis are endowed with distinct developmental potentials and behaviour. For example, the cranial neural crest of gnathostomes gives rise to ectomesenchymal derivatives (for example, the bone and cartilage of the face) in addition to melanocytes, glia and a subset of cranial sensory neurons. By contrast, the trunk neural crest is not able to contribute to cartilage or bone *in vivo*. Rather, these cells form melanocytes, dorsal root and sympathetic ganglia and chromaffin cells. Although the gene regulatory interactions underlying these differences remain unknown, they probably reflect disparities in the mechanisms of specification observed among neural crest subpopulations[33].

Classic heterotopic grafting experiments in the chick demonstrate that the trunk neural crest has restricted developmental potential compared with the cranial population (reviewed in ref. 4). Cranial neural crest cells transplanted to the trunk can not only give rise to all trunk neural crest derivatives, but also form ectopic cartilage nodules that are characteristic of their site of origin[76,77]. By contrast, trunk neural crest transplanted to the head fail to contribute to facial bone and cartilage, although they can form sensory neurons and glia[78]. These results indicate that there are cell-autonomous differences between neural crest subpopulations established during specification. This is consistent with *cis*-regulatory analysis of neural-crest-specifier genes, which shows that expression of both *FoxD3* and *Sox10* in the neural crest is controlled by separate enhancers in the head compared with the trunk[33,34]. Furthermore, activity of these enhancers depends on axial-specific inputs, suggesting that specification of the cranial and trunk neural crest cells relies on different genetic programs[33,38].

The potential of the trunk neural crest has important implications for vertebrate evolution. For instance, it has been suggested that the neural crest played a central part in gnathostome evolution by giving rise to the exoskeleton of early vertebrates such as ostracoderms (armoured fishes)[41]. According to this scenario, at some point during vertebrate evolution the trunk neural crest was endowed with ectomesenchymal potential, which was subsequently lost in extant vertebrates. This hypothesis is based mainly on the fact that the skeletal plates that form the exoskeleton in armoured fishes were composed of dentine, a bona fide neural crest derivative[79,80]. Furthermore, studies in different model organisms suggest that the trunk neural crest exhibits at least some ectomesenchymal potential. For example, fate-map studies in zebrafish and frogs using vital dyes indicate that trunk neural crest contributes to the mesenchyme of the fins[80,81]. Finally, *in vitro* clonal analysis of avian trunk neural crest cells has shown that some clones exhibit gene expression that is characteristic of cartilage and bone[82], suggesting that these cells might possess a latent ectomesenchymal potential, which can be unlocked by environmental signals[83]. These studies suggest that the trunk neural crest might have some residual capacity to form ectomesenchyme, consistent with the hypothesis that the trunk neural crest gave rise to the exoskeleton of basal gnathostomes.

Recently, however, this view has been challenged by a number of studies that employ genetic fate mapping and cell-transplantation analysis to define neural crest contributions in teleost fishes (Box 4). These data show that mesenchyme-derived structures formerly attributed to the trunk neural crest lineage, such as the fin osteoblast, fin mesenchyme and mineral-forming cells of the scales, are in fact of mesodermal origin[84–87]. Taken together, these studies suggest that the trunk neural crest of teleosts has the same developmental restrictions observed in amniotes, calling into question the neural crest origin of the exoskeleton in armoured fishes. Although further studies in other model organisms are necessary for a pan-vertebrate view of trunk neural crest potential, these results indicate that trunk neural crest has been devoid of skeletogenic potential throughout its evolutionary history. These findings suggest that alternative hypotheses for the evolution of the neural crest subpopulations require consideration.

A second scenario is that the cranial neural crest was endowed with gene regulatory mechanisms that are absent from the trunk and may have been 'added on' early in vertebrate evolution. So far, a few developmentally important cranial-specific regulators have been identified. In gnathostomes, for example, *Ets1* (ref. 88) and *Id2* (ref. 89) are enriched in cranial crest cells and are crucial neural-crest-specifier genes for this subpopulation, but their expression is absent from the trunk. This raises the intriguing possibility that the genetic circuits underlying ectomesenchymal potential were added to an ancestral, trunk-like neural crest GRN. According to this view, the ectomesenchymal machinery was either co-opted from the mesoderm[26] or assembled *de novo* in the cranial region. This scenario implies that trunk neural crest cells have a simpler GRN topology than cranial neural crest, an experimentally tractable hypothesis that can be addressed by comparative studies. This view is consistent with the large number of transcriptional regulators that are shared among all neural crest populations, consistent with a common origin.
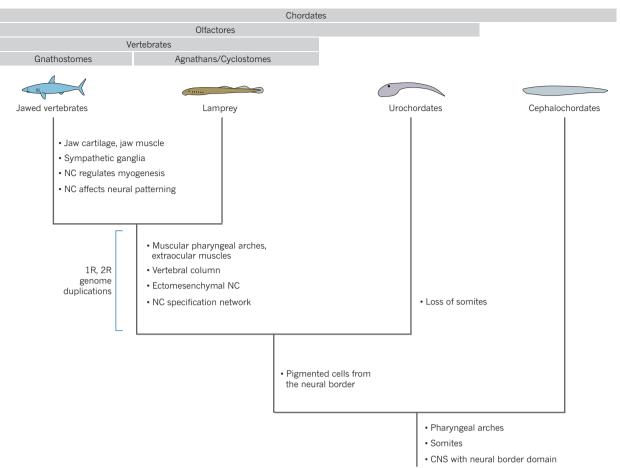
However, a complication is that transcription of genes such as *Sox10* and *FoxD3* is activated uniformly along the entire neural axis, but by distinct enhancers with differential inputs in the trunk compared with cranial regions[33,34]. A third scenario is that neural crest subpopulations may have segregated early in vertebrate evolution and possess different GRN topology. Consistent with enhancer analysis, this hypothesis suggests that many ancestral neural crest GRN connections have been rewired during evolution and that these changes in topology resulted in two populations that have multiple differences in potential and behaviour, despite sharing a similar genetic toolbox. This scenario implies that the trunk and cranial neural crest GRNs have substantial differences, and predicts that pan-neural crest genes are generally controlled by distinct, axial-specific enhancers. Importantly, the hypotheses already discussed can be tested by in-depth analysis of the genetic pathways controlling neural crest formation at different axial levels. In particular, elucidating the circuits controlling ectomesenchymal differentiation of the neural crest will have a great effect on how we interpret the evolution of this cell population. Furthermore, additional neural crest subpopulations exist, including vagal and sacral subtypes, which have distinct migratory pathways and contribute to different derivatives. A more inclusive gene regulatory view of these subpopulations might clarify how the developmental potential of the neural crest

---

BOX 4

Dermal skeleton

A dermal skeleton derived from odontodes is present in many vertebrates, both fossil and living. Dermal skeletal elements among living vertebrates include fin rays (lepidotrichia) of ray-finned (actinoptyerygian) fishes and scales, with multiple subtypes including placoid, ganoid and elasmoid scales in various taxa. Dermal skeletal elements have been proposed to be neural-crest-derived[122] at both cranial and trunk levels. However, recent analyses indicate that osteoblasts responsible for the elasmoid integumentary scales and fin rays of zebrafish derive from mesenchyme of mesodermal origin[87] rather than neural crest[80,123]. Similarly, ossified turtle shells that had been suggested to originate from both mesoderm-derived (endochondral rib) and neural-crest-derived (dermal) osteocytes, instead seem to develop only from mesoderm[124]. These data raise the question of whether the extensive dermal armour of stem gnathostomes originated from mesoderm or neural crest. At trunk levels, these dermal plates may have originated from mesoderm rather than neural crest, although they do arise from neural crest at cranial levels. However, it remains possible that neural crest cells contribute to other scale types, including the placoid scales of cartilaginous fishes that some have argued are more similar to dermal armour of early fishes[87].

Figure 2 | Schematic cladogram of chordate features associated with neural crest cells or their derivatives. Labels at top indicate names of monophyletic groupings below. The timing of duplications is indicated in blue, whereas character changes are indicated by the bullet points. The order of character changes within a stem group is arbitrary. Adapted from ref. 97. CNS, central nervous system; NC, neural crest.

is established at the regulatory level, and have an impact on our views of the evolution of the vertebrate body plan.

Adult neural crest stem cells and post-embryonic growth

Many fossils suggest that the body size of the earliest vertebrates was, like many living invertebrates, quite small[90]. Only later did vertebrates begin to attain larger sizes, presumably through a process that involved extending the duration of post-embryonic growth. Extended growth requires coordinated development of many cell types, possibly including the establishment of stem-cell niches that govern the growth and regeneration of novel tissues.

Until recently, there was little indication of how adult neural crest cell populations were maintained. Recent evidence suggests that amniotes have adult neural crest stem-cell populations that maintain multipotency into adulthood, and which might enable the continuous replenishment of neural-crest-derived tissues[91,92], thus facilitating post-embryonic growth in concert with other tissues. These cells, called Schwann-cell precursors, reside on peripheral nerves and can produce multiple derivatives, including pigment cells and parasympathetic ganglia[93–96]. Whether the GRN underlying differentiation of these neural crest stem cells mirrors that of embryonic progenitor cells is an open and intriguing question that warrants further study. So far, these cells have only been identified in amniotes (in mammals and avians), but there is an obvious need for cells that fill this requirement in other vertebrates, and it is likely that cells such as these originated in early vertebrates.

These studies suggest that the influence of the neural crest in moulding the vertebrate body plan may extend beyond embryonic development, perhaps influencing the increase in size observed in several vertebrate clades. As vertebrates continued to grow post-embryonically, they may

have required the setting aside of a population of neural crest stem cells, in the form of Schwann-cell precursors, that were retained to later stages. The relative proportion of adult tissues that these crest-derived stem cells contribute to is not yet known. Emerging data suggest that this cell population may form many derivatives classically attributed to the embryonic neural crest. Equally, they may represent the key to post-embryonic growth of the vertebrate body and therefore play a heretofore unknown part in promoting vertebrate evolution.

Expansion of neural crest cell types

Development of the neural crest sets vertebrates apart from invertebrate chordates. Formation of this novel cell type was probably facilitated by the addition of a new and uniquely vertebrate 'specification' kernel to the GRN, which in turn conferred multipotency and migratory ability to cells at the neural plate border. During the course of vertebrate evolution, even more derivatives have emerged under the umbrella of the neural crest (for example, additional elements to the peripheral nervous system, elaboration of the jaw or formation of the middle ear). Consolidation of key neural crest specifier genes such as *FoxD3*, *SoxE* and *TFAP2* in the neural-crest-specification module of its GRN may have facilitated evolution of this cell type, by allowing co-option of additional differentiation batteries under the control of neural crest regulators. Arguably, this has made the neural crest one of the most rapidly changing cell types in the vertebrate embryo and has perhaps contributed to the maintenance of neural crest stem cells in adults. ∎

1. Gans, C. & Northcutt, R. G. Neural crest and the origin of vertebrates: a new head. *Science* **220**, 268–273 (1983).

2. Northcutt, R. G. The new head hypothesis revisited. *J. Exp. Zool. B Mol. Dev. Evol.* **304,** 274–297 (2005).
    This article discusses the new head hypothesis in light of more recent data.
3. Gee, H. *Before the Backbone: Views on the Origin of the Vertebrates* (Chapman & Hall, 1996).
4. Le Douarin, N. & Kalcheim, C. *The Neural Crest* (Cambridge Univ. Press, 1999).
5. Patthey, C., Schlosser, G. & Shimeld, S. M. The evolutionary history of vertebrate cranial placodes — I: cell type evolution. *Dev. Biol.* **389,** 82–97 (2014).
6. Schlosser, G., Patthey, C. & Shimeld, S. M. The evolutionary history of vertebrate cranial placodes — II. Evolution of ectodermal patterning. *Dev. Biol.* **389,** 98–119 (2014).
7. Hall, B. K. The neural crest as a fourth germ layer and vertebrates as quadroblastic not triploblastic. *Evol. Dev.* **2,** 3–5 (2000).
8. Sauka-Spengler, T. & Bronner-Fraser, M. Evolution of the neural crest viewed from a gene regulatory perspective. *Genesis* **46,** 673–682 (2008).
9. Holland, N. D. & Chen, J. Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. *Bioessays* **23,** 142–151 (2001).
10. Sauka-Spengler, T., Meulemans, D., Jones, M. & Bronner-Fraser, M. Ancient evolutionary origin of the neural crest gene regulatory network. *Dev. Cell* **13,** 405–420 (2007).
    This work demonstrated that the lamprey has neural crest GRN components that are homologous to those in other vertebrates in both expression pattern and function, indicating that the neural crest GRN is largely shared throughout all vertebrates.
11. Bronner-Fraser, M. & Fraser, S. E. Cell lineage analysis reveals multipotency of some avian neural crest cells. *Nature* **335,** 161–164 (1988).
12. Bronner-Fraser, M. & Fraser, S. Developmental potential of avian trunk neural crest cells *in situ. Neuron* **3,** 755–766 (1989).
13. Frank, E. & Sanes, J. R. Lineage of neurons and glia in chick dorsal root ganglia: analysis *in vivo* with a recombinant retrovirus. *Development* **111,** 895–908 (1991).
14. Dupin, E., Calloni, G. W. & Le Douarin, N. M. The cephalic neural crest of amniote vertebrates is composed of a large majority of precursors endowed with neural, melanocytic, chondrogenic and osteogenic potentialities. *Cell Cycle* **9,** 238–249 (2010).
15. Calloni, G. W., Le Douarin, N. M. & Dupin, E. High frequency of cephalic neural crest cells shows coexistence of neurogenic, melanogenic, and osteogenic differentiation capacities. *Proc. Natl Acad. Sci. USA* **106,** 8947–8952 (2009).
16. Heimberg, A. M., Cowper-Sal·lari, R., Sémon, M., Donoghue, P. C. J. & Peterson, K. J. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc. Natl Acad. Sci. USA* **107,** 19379–19383 (2010).
17. Donoghue, P. C. J. & Keating, J. N. Early vertebrate evolution. *Palaeontology* **57,** 879–893 (2014).
18. Oisi, Y., Ota, K. G., Kuraku, S., Fujimoto, S. & Kuratani, S. Craniofacial development of hagfishes and the evolution of vertebrates. *Nature* **493,** 175–180 (2013).
19. Ota, K. G. & Kuratani, S. Cyclostome embryology and early evolutionary history of vertebrates. *Integr. Comp. Biol.* **47,** 329–337 (2007).
20. Shimeld, S. M. & Donoghue, P. C. J. Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). *Development* **139,** 2091–2099 (2012).
21. Hall, B. K. & Gillis, J. A. Incremental evolution of the neural crest, neural crest cells and neural crest-derived skeletal tissue. *J. Anat.* **222,** 19–31 (2013).
22. Mongera, A. *et al.* Genetic lineage labeling in zebrafish uncovers novel neural crest contributions to the head, including gill pillar cells. *Development* **140,** 916–925 (2013).
    This paper identifies gill pillar cells, which are crucial for gill structure throughout vertebrates, as neural crest derivatives.
23. Häming, D. *et al.* Expression of sympathetic nervous system genes in lamprey suggests their recruitment for specification of a new vertebrate feature. *PLoS ONE* **6,** e26543 (2011).
24. Medeiros, D. M. The evolution of the neural crest: new perspectives from lamprey and invertebrate neural crest-like cells. *Wiley Interdiscip. Rev. Dev. Biol.* **2,** 1–15 (2013).
25. Meulemans, D. & Bronner-Fraser, M. Central role of gene cooption in neural crest evolution. *J. Exp. Zool. B. Mol. Dev. Evol.* **304,** 298–303 (2005).
26. Jandzik, D. *et al.* Evolution of the new vertebrate head by co-option of an ancient chordate skeletal tissue. *Nature* **518,** 534–537 (2015).
    This crucial paper identifies cellular cartilage in a cephalochordate, lending support to the contention that neural-crest-derived cartilage was co-opted from other tissues rather than constructed *de novo*.
27. Davidson, E. H. *The Regulatory Genome* (Academic, 2010).
28. Erwin, D. H. & Davidson, E. H. The evolution of hierarchical gene regulatory networks. *Nature Rev. Genet.* **10,** 141–148 (2009).
29. de Crozé, N., Maczkowiak, F. & Monsoro-Burq, A. H. Reiterative AP2a activity controls sequential steps in the neural crest gene regulatory network. *Proc. Natl Acad. Sci. USA* **108,** 155–160 (2011).
30. Nieto, M. A., Sargent, M. G., Wilkinson, D. G. & Cooke, J. Control of cell behavior during vertebrate development by *Slug*, a zinc finger gene. *Science* **264,** 835–839 (1994).
31. Labosky, P. A. & Kaestner, K. H. The winged helix transcription factor Hfh2 is expressed in neural crest and spinal cord during mouse development. *Mech. Dev.* **76,** 185–190 (1998).
32. Dottori, M., Gross, M. K., Labosky, P. & Goulding, M. The winged-helix transcription factor Foxd3 suppresses interneuron differentiation and promotes neural crest cell fate. *Development* **128,** 4127–4138 (2001).
33. Simões-Costa, M. S., McKeown, S. J., Tan-Cabugao, J., Sauka-Spengler, T. & Bronner, M. E. Dynamic and differential regulation of stem cell factor FoxD3 in the neural crest is encrypted in the genome. *PLoS Genet.* **8,** e1003142 (2012).
34. Betancur, P., Bronner-Fraser, M. & Sauka-Spengler, T. Genomic code for Sox10 activation reveals a key regulatory enhancer for cranial neural crest. *Proc. Natl Acad. Sci. USA* **107,** 3570–3575 (2010).
35. McKeown, S. J., Lee, V. M., Bronner-Fraser, M., Newgreen, D. F. & Farlie, P. G. Sox10 overexpression induces neural crest-like cells from all dorsoventral levels of the neural tube but inhibits differentiation. *Dev. Dyn.* **233,** 430–444 (2005).
36. Meulemans, D. & Bronner-Fraser, M. Gene-regulatory interactions in neural crest evolution and development. *Dev. Cell* **7,** 291–299 (2004).
37. Betancur, P., Bronner-Fraser, M. & Sauka-Spengler, T. Assembling neural crest regulatory circuits into a gene regulatory network. *Annu. Rev. Cell Dev. Biol.* **26,** 581–603 (2010).
38. Simões-Costa, M. & Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe. *Development* **142,** 242–257 (2015).
39. Sauka-Spengler, T. & Bronner-Fraser, M. A gene regulatory network orchestrates neural crest formation. *Nature Rev. Mol. Cell Biol.* **9,** 557–568 (2008).
40. Simões-Costa, M. & Bronner, M. E. Insights into neural crest development and evolution from genomic analysis. *Genome Res.* **23,** 1069–1080 (2013).
41. Donoghue, P. C. J., Graham, A. & Kelsh, R. N. The origin and evolution of the neural crest. *Bioessays* **30,** 530–541 (2008).
42. Meulemans, D. & Bronner-Fraser, M. Amphioxus and lamprey AP-2 genes: implications for neural crest evolution and migration patterns. *Development* **129,** 4953–4962 (2002).
43. Ota, K. G., Kuraku, S. & Kuratani, S. Hagfish embryology with reference to the evolution of the neural crest. *Nature* **446,** 672–675 (2007).
44. Aybar, M. J., Nieto, M. A. & Mayor, R. *Snail* precedes *Slug* in the genetic cascade required for the specification and migration of the *Xenopus* neural crest. *Development* **130,** 483–494 (2003).
45. LaBonne, C. & Bronner-Fraser, M. Snail-related transcriptional repressors are required in *Xenopus* for both the induction of the neural crest and its subsequent migration. *Dev. Biol.* **221,** 195–205 (2000).
46. Oram, K. F. & Gridley, T. Mutations in Snail family genes enhance craniosynostosis of *Twist1* haplo-insufficient mice: implications for Saethre-Chotzen Syndrome. *Genetics* **170,** 971–974 (2005).
47. Rogers, C. D., Saxena, A. & Bronner, M. E. Sip1 mediates an E-cadherin-to-N-cadherin switch during cranial neural crest EMT. *J. Cell Biol.* **203,** 835–847 (2013).
48. Jeffery, W. R. *et al.* Trunk lateral cells are neural crest-like cells in the ascidian *Ciona intestinalis*: insights into the ancestry and evolution of the neural crest. *Dev. Biol.* **324,** 152–160 (2008).
49. Abitua, P. B., Wagner, E., Navarrete, I. A. & Levine, M. Identification of a rudimentary neural crest in a non-vertebrate chordate. *Nature* **492,** 104–107 (2012).
    This paper argues that a gene regulatory network acting in the *C. intestinalis* a9.49 cell lineage is homologous to the GRN of vertebrate neural crest, and suggests that co-option of mesenchymal migration controls might have facilitated expansion of neural crest derivatives in early vertebrates.
50. Ivashkin, E. & Adameyko, I. Progenitors of the protochordate ocellus as an evolutionary origin of the neural crest. *Evodevo* **4,** 12 (2013).
51. Yu, J.-K., Meulemans, D., McKeown, S. J. & Bronner-Fraser, M. Insights from the amphioxus genome on the origin of vertebrate neural crest. *Genome Res.* **18,** 1127–1132 (2008).
    This paper showed that an AmphiFoxD regulatory element was capable of driving expression in chicken somites, but not in neural crest, suggesting that novel regulatory elements were required for AmphiFoxD incorporation into the neural crest GRN.
52. Langeland, J. A., Tomsa, J. M., Jackman, W. R. & Kimmel, C. B. An amphioxus snail gene: expression in paraxial mesoderm and neural plate suggests a conserved role in patterning the chordate embryo. *Dev. Genes Evol.* **208,** 569–577 (1998).
53. Denes, A. S. *et al.* Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell* **129,** 277–288 (2007).
54. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
55. Yu, J.-K., Holland, N. D. & Holland, L. Z. An amphioxus winged helix/forkhead gene, *AmphiFoxD*: insights into vertebrate neural crest evolution. *Dev. Dyn.* **225,** 289–297 (2002).
56. Yu, J.-K. S. The evolutionary origin of the vertebrate neural crest and its developmental gene regulatory network — insights from amphioxus. *Zoology (Jena)* **113,** 1–9 (2010).
57. Parker, H. J., Bronner, M. E. & Krumlauf, R. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. *Nature* **514,** 490–493 (2014).
58. Dehal, P. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298,** 2157–2167 (2002).
59. Holland, L. Z. *et al.* The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18,** 1100–1111 (2008).
60. Yu, J.-K., Holland, N. D. & Holland, L. Z. Tissue-specific expression of FoxD reporter constructs in amphioxus embryos. *Dev. Biol.* **274,** 452–461 (2004).
61. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144,** 970–985 (2011).
62. Ono, H., Kozmik, Z., Yu, J.-K. & Wada, H. A novel N-terminal motif is responsible for the evolution of neural crest-specific gene-regulatory activity in vertebrate FoxD3. *Dev. Biol.* **385,** 396–404 (2014).

63. Taylor, J. S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* **38**, 615–643 (2004).

64. Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A. & Van de Peer, Y. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl Acad. Sci. USA* **101**, 1638–1643 (2004).

65. Crow, K. D., Wagner, G. P. & SMBE Tri-National Young Investigators. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? *Mol. Biol. Evol.* **23**, 887–892 (2006).

66. Holland, P. W., Garcia-Fernàndez, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Dev. Suppl.* **1994,** 125–133 (1994).

67. Holland, L. Z. Evolution of new characters after whole genome duplications: insights from amphioxus. *Semin. Cell Dev. Biol.* **24,** 101–109 (2013).

68. Smith, J. J. *et al.* Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genet.* **45,** 415–421 (2013).

69. Kuraku, S. Insights into cyclostome phylogenomics: pre-2R or post-2R. *Zoolog. Sci.* **25,** 960–968 (2008).

70. Kuraku, S., Meyer, A. & Kuratani, S. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Mol. Biol. Evol.* **26,** 47–59 (2009).

71. Smith, J. J. The sea lamprey meiotic map resolves ancient vertebrate genome duplications. Preprint at http://dx.doi.org/10.1101/008953 (2014).

72. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol.* **3,** e245 (2005).

73. Kassahn, K. S., Dang, V. T., Wilkins, S. J., Perkins, A. C. & Ragan, M. A. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* **19,** 1404–1418 (2009).

74. Kim, Y. J. *et al.* Generation of multipotent induced neural crest by direct reprogramming of human postnatal fibroblasts with a single transcription factor. *Cell Stem Cell* **15,** 497–506 (2014).

75. Emes, R. D. *et al.* Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nature Neurosci.* **11,** 799–806 (2008).

76. Le Douarin, N. M. & Teillet, M. A. Experimental analysis of the migration and differentiation of neuroblasts of the autonomic nervous system and of neuroectodermal mesenchymal derivatives, using a biological cell marking technique. *Dev. Biol.* **41,** 162–184 (1974).

77. Le Lièvre, C. S., Schweizer, G. G., Ziller, C. M. & Le Douarin, N. M. Restrictions of developmental capabilities in neural crest cell derivatives as tested by *in vivo* transplantation experiments. *Dev. Biol.* **77,** 362–378 (1980).

78. Le Lièvre, C. S. & Le Douarin, N. M. Mesenchymal derivatives of the neural crest: analysis of chimaeric quail and chick embryos. *J. Embryol. Exp. Morphol.* **34,** 125–154 (1975).

79. Sire, J.-Y., Donoghue, P. C. J. & Vickaryous, M. K. Origin and evolution of the integumentary skeleton in non-tetrapod vertebrates. *J. Anat.* **214,** 409–440 (2009).

80. Smith, M., Hickman, A., Amanze, D., Lumsden, A. & Thorogood, P. Trunk neural crest origin of caudal fin mesenchyme in the zebrafish *Brachydanio rerio. Proc. R. Soc. Lond. B* **256,** 137–145 (1994).

81. Collazo, A., Bronner-Fraser, M. & Fraser, S. E. Vital dye labelling of *Xenopus laevis* trunk neural crest reveals multipotency and novel pathways of migration. *Development* **118,** 363–376 (1993).

82. Coelho-Aguiar, J. M., Le Douarin, N. M. & Dupin, E. Environmental factors unveil dormant developmental capacities in multipotent progenitors of the trunk neural crest. *Dev. Biol.* **384,** 13–25 (2013).

83. McGonnell, I. M. & Graham, A. Trunk neural crest has skeletogenic potential. *Curr. Biol.* **12,** 767–771 (2002).

84. Lee, R. T. H., Knapik, E. W., Thiery, J. P. & Carney, T. J. An exclusively mesodermal origin of fin mesenchyme demonstrates that zebrafish trunk neural crest does not generate ectomesenchyme. *Development* **140,** 2923–2932 (2013).
This recent paper finds that neural crest cells at trunk levels do not contribute to fin mesenchyme, in contrast to earlier claims.

85. Lee, R. T. H., Thiery, J. P. & Carney, T. J. Dermal fin rays and scales derive from mesoderm, not neural crest. *Curr. Biol.* **23,** R336–R337 (2013).

86. Shimada, A. *et al.* Trunk exoskeleton in teleosts is mesodermal in origin. *Nature Commun.* **4,** 1639 (2013).
This paper suggests that mesoderm, and not neural crest, gives rise to the exoskeleton at trunk levels, conflicting with findings of earlier studies and calling into question whether trunk neural crest has ectomesenchymal capability.

87. Mongera, A. & Nüsslein-Volhard, C. Scales of fish arise from mesoderm. *Curr. Biol.* **23,** R338–R339 (2013).

88. Théveneau, E., Duband, J.-L. & Altabef, M. Ets-1 confers cranial features on neural crest delamination. *PLoS ONE* **2,** e1142 (2007).

89. Martinsen, B. J. & Bronner-Fraser, M. Neural crest specification regulated by the helix-loop-helix repressor Id2. *Science* **281,** 988–991 (1998).

90. Janvier, P. *Early Vertebrates* (Oxford Univ. Press, 1996).

91. Jinno, H. *et al.* Convergent genesis of an adult neural crest-like dermal stem cell from distinct developmental origins. *Stem Cells* **28,** 2027–2040 (2010).

92. Morrison, S. J., White, P. M., Zock, C. & Anderson, D. J. Prospective identification, isolation by flow cytometry, and *in vivo* self-renewal of multipotent mammalian neural crest stem cells. *Cell* **96,** 737–749 (1999).

93. Dyachuk, V. *et al.* Neurodevelopment. Parasympathetic neurons originate from nerve-associated peripheral glial progenitors. *Science* **345,** 82–87 (2014).
One of two papers demonstrating that Schwann-cell precursors also give rise to parasympathetic neurons.

94. Espinosa-Medina, I. *et al.* Neurodevelopment. Parasympathetic ganglia derive from Schwann cell precursors. *Science* **345,** 87–90 (2014).
The second of two papers demonstrating that Schwann-cell precursors are a source of parasympathetic neurons.

95. Adameyko, I. *et al.* Schwann cell precursors from nerve innervation are a cellular origin of melanocytes in skin. *Cell* **139,** 366–379 (2009).
This paper identifies Schwann-cell precursors as a major source of pigment cells in chicken and mouse.

96. Krause, M. P. *et al.* Direct genesis of functional rodent and human Schwann cells from skin mesenchymal precursors. *Stem Cell Rep.* **3,** 85–100 (2014).

97. Green, S. A. & Bronner, M. E. The lamprey: a jawless vertebrate model system for examining origin of the neural crest and other vertebrate traits. *Differentiation* **87,** 44–51 (2014).

98. Northcutt, R. G. & Gans, C. The genesis of neural crest and epidermal placodes: a reinterpretation of vertebrate origins. *Q. Rev. Biol.* **58,** 1–28 (1983).

99. Graham, A. Deconstructing the pharyngeal metamere. *J. Exp. Zool. B Mol. Dev. Evol.* **310,** 336–344 (2008).

100. McCauley, D. W. & Bronner-Fraser, M. Neural crest contributions to the lamprey head. *Development* **130,** 2317–2327 (2003).

101. Gillis, J. A., Fritzenwanker, J. H. & Lowe, C. J. A stem-deuterostome origin of the vertebrate pharyngeal transcriptional network. *Proc. Biol. Sci.* **279,** 237–246 (2012).

102. Smith, A. B. The pre-radial history of echinoderms. *Geolog. J.* **40,** 255–280 (2005).

103. Graham, A. & Richardson, J. Developmental and evolutionary origins of the pharyngeal apparatus. *Evodevo* **3,** 24 (2012).

104. Yasui, K., Kaji, T., Morov, A. R. & Yonemura, S. Development of oral and branchial muscles in lancelet larvae of *Branchiostoma japonicum. J. Morphol.* **275,** 465–477 (2014).

105. Trinajstic, K. *et al.* Fossil musculature of the most primitive jawed vertebrates. *Science* **341,** 160–164 (2013).

106. Matsuoka, T. *et al.* Neural crest origins of the neck and shoulder. *Nature* **436,** 347–355 (2005).

107. Köntges, G. & Lumsden, A. Rhombencephalic neural crest segmentation is preserved throughout craniofacial ontogeny. *Development* **122,** 3229–3242 (1996).

108. Fraser, G. J., Cerny, R., Soukup, V., Bronner-Fraser, M. & Streelman, J. T. The odontode explosion: the origin of tooth-like structures in vertebrates. *Bioessays* **32,** 808–817 (2010).

109. Murdock, D. J. E. *et al.* The origin of conodonts and of vertebrate mineralized skeletons. *Nature* **502,** 546–549 (2013).

110. Janvier, P. Inside-out turned upside-down. *Nature* **502,** 457–458 (2013).

111. Creuzet, S. E., Martinez, S. & Le Douarin, N. M. The cephalic neural crest exerts a critical effect on forebrain and midbrain development. *Proc. Natl Acad. Sci. USA* **103,** 14033–14038 (2006).

112. Le Douarin, N. M., Couly, G. & Creuzet, S. E. The neural crest is a powerful regulator of pre-otic brain development. *Dev. Biol.* **366,** 74–82 (2012).

113. Aguiar, D. P., Sghari, S. & Creuzet, S. The facial neural crest controls fore- and midbrain patterning by regulating Foxg1 expression through Smad1 activity. *Development* **141,** 2494–2505 (2014).

114. Holland, L. Z. *et al.* Evolution of bilaterian central nervous systems: a single origin? *Evodevo* **4,** 27 (2013).

115. Pani, A. M. *et al.* Ancient deuterostome origins of vertebrate brain signalling centres. *Nature* **483,** 289–294 (2012).

116. Noden, D. M. & West, P. F. The differentiation and morphogenesis of craniofacial muscles. *Dev. Dyn.* **235,** 1194–1218 (2006).

117. Sambasivan, R., Kuratani, S. & Tajbakhsh, S. An eye on the head: the development and evolution of craniofacial muscles. *Development* **138,** 2401–2415 (2011).

118. Lee, G.-H., Chang, M.-Y., Hsu, C.-H. & Chen, Y.-H. Essential roles of basic helix-loop-helix transcription factors, Capsulin and Musculin, during craniofacial myogenesis of zebrafish. *Cell. Mol. Life Sci.* **68,** 4065–4078 (2011).

119. Tzahor, E. Heart and craniofacial muscle development: a new developmental theme of distinct myogenic fields. *Dev. Biol.* **327,** 273–279 (2009).

120. Kelly, R. G. Core issues in craniofacial myogenesis. *Exp. Cell Res.* **316,** 3034–3041 (2010).

121. Johnels, A. G. On the peripheral autonomic nervous system of the trunk region of *Lampetra planeri. Acta Zool.* **37,** 251–286 (1956).

122. Donoghue, P. C. J. & Sansom, I. J. Origin and early evolution of vertebrate skeletonization. *Microsc. Res. Tech.* **59,** 352–372 (2002).

123. Kague, E. *et al.* Skeletogenic fate of zebrafish cranial and trunk neural crest. *PLoS ONE* **7,** e47394 (2012).

124. Hirasawa, T., Nagashima, H. & Kuratani, S. The endoskeletal origin of the turtle carapace. *Nature Commun.* **4,** 2107 (2013).

Facts and fancies about early fossil chordates and vertebrates

Philippe Janvier[1]

The interrelationships between major living vertebrate, and even chordate, groups are now reasonably well resolved thanks to a large amount of generally congruent data derived from molecular sequences, anatomy and physiology. But fossils provide unexpected combinations of characters that help us to understand how the anatomy of modern groups was progressively shaped over millions of years. The dawn of vertebrates is documented by fossils that are preserved as either soft–tissue imprints, or minute skeletal fragments, and it is sometimes difficult for palaeontologists to tell which of them are reliable vertebrate remains and which merely reflect our idea of an ancestral vertebrate.

Vertebrates are a very small group among animals, but they show, along with arthropods and possibly echinoderms, a large number of 'fossilizable' complex characters that can be analysed to reconstruct their relationships; however, most of their anatomically informative fossil record appeared relatively late, about 470 million years ago (Ma). During the past 20 years or so, the fossil record of Palaeozoic era, 535–250 million year (Myr) old, jawless vertebrates has been enriched by the discovery of spectacular soft-bodied fossils preserved as imprints in famous fossil sites such as Chengjiang (535 Myr old) in China[1] and the Burgess Shale in Canada[2] (510 Myr old), but also in other, younger sites that yield exceptional preservation of soft tissues (referred to as 'Konservat-Lagerstätte'). These fossils, long considered to be trivial by palaeoanatomists, have gained a new dimension thanks to investigation and imaging techniques that allow the actual nature of the preserved tissues to be identified, as well as a better understanding of the processes involved in decay and fossilization, thereby avoiding their overinterpretation[3–6].

Palaeontologists have been extensively tracing the earliest evidence for typical vertebrate hard tissues, such as bone, calcified cartilage, dentine (the 'ivory' of our teeth) or enamel, generally in the form of bone fragments, isolated scales or denticles made up of bioapatite (calcium phosphate) and found scattered in early Palaeozoic sediments[7,8]. This search for vertebrate 'microremains' or 'ichthyoliths' (often the only available vertebrate remains in the early Palaeozoic) yielded a large diversity of skeletal elements that could be compared with those of previously known, younger, complete fossils that belong to the major vertebrate groups, and provided evidence for the antiquity of most classic vertebrate hard tissues at least since the Lower or Middle Ordovician (about 477 Ma). However, this research also yielded some skeletal elements that, although suggesting the shape of scales or teeth, do not show all the characteristics of hitherto recognized vertebrate hard tissues. Such cases are frequent among Ordovican to Silurian (480–420 Myr old) microremains, which are dismissed by some, but regarded as possible vertebrates by others. The vertebrate fossil record is documented by an abundance of articulated specimens from periods since the late Silurian (about 430 Ma), but is either poorly represented or very puzzling in earlier periods. However, late Silurian (430 Myr old) articulated vertebrates still turn up (in Scotland, Canada and China[9,10]), and hint at exciting issues in deeper vertebrate history.

This may give the reader the impression that the early history (before the late Silurian) of vertebrate evolution is documented by fossils that look rather like squashed slugs and crushed lobster carapaces, although sometimes articulated. Uninformative data indeed, but, practically, it is all we

can offer, except for extremely rare three-dimensionally preserved jawless vertebrates, such as the Ordovician astraspids and arandaspids[11–13], which document the first occurrence of an extensive exoskeleton (or dermal skeleton, the superficial skeleton of vertebrates) with site-specific bones and a lateral-line system (the superficial sense organ of fishes).

Living vertebrates fall into two major clades, the cyclostomes (hagfishes and lampreys) and the gnathostomes (jawed vertebrates). Only the latter produce bone and dentine. Therefore, current vertebrate phylogenies that include fossils suggest that all the Palaeozoic jawless vertebrates that display at least an exoskeleton are more closely related to gnathostomes than to cyclostomes, and are thus 'stem gnathostomes', although lacking jaws[14]. These jawless stem gnathostomes that possess a calcified skeleton are informally referred to as 'ostracoderms' for historical reasons, but form
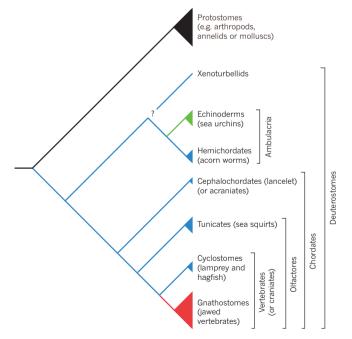
Figure 1 | Interrelationships of the major extant deuterostome clades. Distribution of the major tissues potentially preserved in fossil deuterostomes: no calcified hard tissue except for occasional calcified cartilage in vertebrates (blue), calcitic skeleton (green) and bone, dentine, enamel or enameloid (red).

[1]Muséum National d'Histoire Naturelle, UMR7207 du CNRS (Sorbonne Universités, UPMC), 8 rue Buffon, 75231 Paris, France.
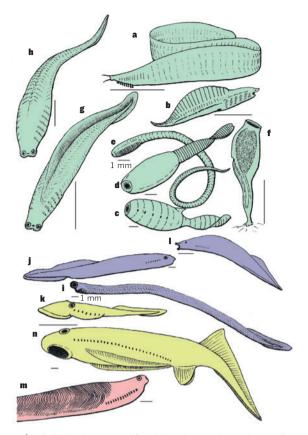
Figure 2 | Soft-bodied presumed fossil chordates and vertebrates, from the Cambrian (green), Silurian (pink), Devonian (yellow) and Carboniferous (purple) periods. **a**, *Pikaia* was long regarded as a chordate, but is now considered to be either of uncertain affinity, or possibly a close relative of yunnanozoans (adapted from ref. 22). **b**, The yunnanozoan *Haikouella* is a possible stem deuterostome or stem vertebrate (adapted from ref. 28). **c, d**, The vetulicolans *Didazoon* (**c**) and *Banffia* (**d**) are possible stem chordates, stem deuterostomes or stem protostomes (adapted from refs 31, 33). **e**, *Cathaymyrus* is a possible stem cephalochordate (adapted from ref. 34). **f**, *Shankouclava* is a likely tunicate (adapted from ref. 36). **g, h**, *Haikouichthys* (**g**) and *Metaspriggina* (**h**) are stem vertebrates (based on refs 40, 41). **i**, *Clydagnathus* is a euconodont (adapted from ref. 46). **j, k**, *Mayomyzon* (**j**) and *Priscomyzon* (**k**) are two fossil lampreys (adapted from refs 15, 17). **l**, *Myxinikela* is a probable hagfish (adapted from ref. 19). **m**, *Jamoytius* is a jawless stem gnathostome with thin mineralized body scales (adapted from ref. 60). **n**, *Euphanerops*, a jawless vertebrate whose calcified cartilage displays a lamprey-like annular cartilage and branchial basket (adapted from refs 61, 64). Scale bars are 10 mm (**a**–**d**, **f**–**h**, **j**–**n**) and 1 mm (**e**, **i**).

a grade: an array of groups that are more and more closely related to jawed vertebrates and whose anatomy documents the progressive assembly of the gnathostome body plan before the rise of jaws. By contrast, there is no evidence that cyclostomes have ever produced a mineralized skeleton, and neither the four fossil lampreys[15–18], nor the two possible fossil hagfishes[19,20] show any clear indication of a mineralized skeleton.

Soft-bodied chordates and wishful thinking

The bestiary of the Chengjiang and Burgess Shale sites[1,2] comprises a number of animals that have been referred to as either chordates or other deuterostome groups (Fig. 1). Most of these fossils have been referred to as chordates because they show at least some indication of either a notochord (the axial support of chordates, and precursor of the vertebral column), a segmented body structure or gill slits. Although the segmentation of the body musculature and gill apparatus has different developmental causes[21], it is often regarded as a 'signature' of the chordates, but is readily distinguished from the metamery (repeated parts) of arthropods or annelids. Notably, this was the case for *Pikaia* (Fig. 2a), from the Burgess Shale, whose body shows indications of a series of myomeres (muscle blocks)

and a notochord, but whose head bears peculiar appendages (regarded as respiratory organs) and tentacles that are at odds with vertebrate anatomy[22]. Despite the exquisite preservation of numerous specimens of *Pikaia*, this long iconic 'vertebrate ancestor'[23] remains an enigma, and opinions about its affinities oscillate between the chordate hypothesis and a convergent morphology in some protostomes (the sister group of deuterostomes)[22] (Fig. 1). Yunnanozoans (*Yunnanozoon* and *Haikouella*; Fig. 2b) from Chengjiang have also been referred to as chordates[24] because of their presumed notochord, segmented body musculature covered by a cuticle and their seemingly vertebrate-like series of six gill pairs. Notably, they have been referred to as either stem deuterostomes[25], hemichordates, cephalochordates or stem vertebrates[26–28]. The controversy between the advocates of the stem-vertebrate[29] and stem-deuterostome[30] hypotheses reflects the difficulty in assessing the nature of the actual tissues and anatomical characters observed in these fossils. Vetulicolans[31,32] (*Vetulicola*, *Xidazoon*, *Didazoon* and *Pomatrum*; Fig. 2c) from Chengjiang and the somewhat similar *Banffia* (Fig. 2d) from the Burgess Shale display a bipartite structure, with a balloon-shaped, cuticle-covered head laterally pierced by five presumed gill openings, and a flattened segmented tail[33]. *Banffia*, however, seems devoid of gill openings and displays midgut diverticulae that rather suggest a protostome anatomy[33]. Again, the vetulicolan's gill openings might suggest a stem deuterostome, but the purported presence of an endostyle (a gland unique to chordates) suggests stem chordate affinity[32]. *Cathaymyrus* (Fig. 2e), from Chengjiang, was described as "*Pikaia*-like"[34]. It has a worm-shaped body with a long series of myomeres, and a distinct row of closely set pharyngeal slits that resemble those of cephalochordates. Other presumed chordates from Chengjiang are the debated tunicates *Cheungkongella*[35] and *Shankouclava*[36] (Fig. 2f). As a whole, all these presumed chordates from the Cambrian, mostly preserved as soft-tissue imprints, only provide tenuous information about their possible phylogenetic relationships. And, despite their often spectacular preservation, there is a risk of overinterpreting their anatomy in the light of widely different living organisms. A notable example of this problem is *Ainiktozoon* (a much younger fossil from the Silurian (430 Ma) of Scotland), which has been interpreted both as a possible chordate because of its segmented body[37] and as a thylacocephalan — a peculiar extinct arthropod group[38].

The myllokunmingiids (for example, *Myllokunmingia* and *Haikouichthys*; Fig. 2g)[39,40] from Chengjiang and the similar *Metaspriggina*[41] (Fig. 2h) from the Burgess Shale look more familiar to vertebrate specialists, as they are clearly 'fish-like'. Despite their similarities, *Metaspriggina* provides better information about the arrangement of gill bars and eye structure. Although only a small number of characters can actually be observed on this kind of material, character analyses have resolved myllokunmingiids as paraphyletic, with *Myllokunmingia* as a stem vertebrate, and *Haikouichthys* as a stem lamprey[39]. More recent analyses suggest that all myllokunmingiids, and probably *Metaspriggina*, are stem vertebrates, but appear in a basal polytomy in the vertebrate tree, more crownward than *Pikaia*, but less so than any crown-group vertebrate (the last common ancestor to living vertebrates and all their fossil relatives)[41]. By combining myllokunmingiids and *Metaspriggina* data, a better reconstruction of the most likely Cambrian vertebrates is possible — a jawless 'fish' with a pair of large, anterodorsally facing camera eyes, a small median olfactory organ, 5–7 pairs of gill arches, a stomach, a series of chevron-shaped myomeres and a median fin web (Fig. 2g, h), thereby remotely resembling old hypothetical reconstructions of ancestral vertebrates[42] (Box 1).

The soft-bodied fossil record of the vertebrates is not limited to the Cambrian, and after the Cambrian 'squashed slug' episode comes the saga of the conodonts. Conodonts are minute tooth- or comb-like elements, or denticles, that are made up of bioapatite (like vertebrate teeth) and occur in marine sediments from the Cambrian to the Late Triassic (about 530–200 Ma). Depending on their internal structure, conodonts fall into three groups: protoconodonts, paraconodonts and euconodonts, the latter being the only monophyletic one[43]. For more than a century, conodonts have received diverse, sometimes fanciful interpretations, until the 1983 publication of the first 'conodont-bearing animal', from the Carboniferous

(330 Ma): a conodont assemblage located in the mouth of an eel-shaped animal preserved as a soft-tissue imprint[44]. Other specimens have since turned up[45], but so far all known articulated conodont-bearing animals are euconodonts. Anatomically, a euconodont-bearing animal has a small head with large paired eyes, a mouth or pharynx containing a large number of denticles, an elongated eel-shaped body with chevron-shaped myomeres, and a small caudal fin supported by possibly cartilaginous rods (Fig. 2i)[46,47]. Superficially, this agrees with vertebrate morphology, although the absence of more typical vertebrate structures, such as gill arches, remains puzzling. The most contentious question was whether euconodont denticle tissues were homologous with vertebrate teeth and odontodes (skin denticles; Fig. 3a), a scenario that was advocated by some[48], but rejected by others[49]. This controversy was finally resolved with the demonstration, by means of high-resolution microtomographic techniques, that euconodont denticle structure and growth were largely at odds with that of vertebrate odontodes[43]. Nevertheless, there remains a chordate- or vertebrate-like aspect to the euconodont body imprints, which does not preclude their position as either stem vertebrates or stem cyclostomes (Fig. 4). During the past 15 years, euconodonts were almost constantly considered in phylogenetic analyses of early vertebrates, and their position as basal-most stem gnathostomes was essentially supported by the presence of the phosphatic denticles[50], which were then assumed to be homologues of gnathostome hard tissues, but lacking in all cyclostomes. However, an old hypothesis that euconodonts might be allied to cyclostomes, and more specifically hagfishes, periodically reappears in the literature[51–53]. For example, the enigmatic Carboniferous protoconodont-like soft-bodied fossil *Conopiscius*[54] shows, like euconodont-bearing animals, a series of chevron-shaped myomeres, but a single pair of hollow, weakly mineralized denticles[52]. It has been suggested that conodont denticles were partly or entirely capped with a keratinous tissue[51,52], which would remain in living cyclostomes. This hypothesis has now been dismissed[55]. The controversy about the homology of the para- and euconodont elements now seems to be settled, and all that soft-tissue data can currently suggest is that euconodonts might be either stem vertebrates, stem cyclostomes (Fig. 4) or, less likely, stem lampreys or stem hagfishes.

Other possible soft-bodied fossil chordates occur here and there, notably in Silurian to Carboniferous rocks, and some are more readily recognized as vertebrates, because they superficially resemble living hagfishes or lampreys. However, the risk of being misled by wishful thinking when making such comparisons is much the same as with odd Cambrian fossils. The fossil lampreys came as a surprise when first discovered in Carboniferous 300-Myr-old rocks, because of their striking overall resemblance to modern forms. *Mayomyzon*[15] (Fig. 2j), preserved as an imprint from the Mazon Creek Lagerstätte in Illinois, looks somewhat like a radiograph of a small modern lamprey. The image shows the outline of the body, the gill pouches and the characteristic cartilages of the 'tongue' apparatus. Other fossil lampreys turned up in the Carboniferous[16] and the Late Devonian (around 360 Ma)[17]. The latter, *Priscomyzon* (Fig. 2k), shows annular cartilage that supports the characteristic oral funnel. The two presumed fossil hagfishes, both coeval with *Mayomyzon*, are more questionable. *Myxinikela*[19] (Fig. 2l) has cartilage imprints and tentacles that do resemble those of hagfishes, but *Myxineidus*[20] was referred to as a hagfish based only on the impression of two V-shaped rows of keratinous teeth that resemble those of living hagfishes. The Mazon Creek Lagerstätte has also yielded peculiar presumed soft-bodied jawless vertebrates, *Pipiscius* and *Gilpichthys*[56]. The former has a lamprey-like oral funnel, and the latter shows possible impressions of sharp, non-mineralized teeth that resemble those of hagfishes[57]. Yet this interpretation remains controversial[58].

Another peculiar Palaeozoic soft-bodied vertebrate is *Jamoytius* (Fig. 2m)[59], from the Silurian (about 438 Ma) of Scotland, which was first regarded as an 'ancestral chordate'. New investigations show that the series of W-shaped imprints on the trunk of *Jamoytius* are not merely soft-tissue imprints of myomeres, but weakly mineralized scales[60]. With its median nostril and about ten gill openings, *Jamoytius* is otherwise suggestive of a lamprey and is often regarded as closely related to the younger, Devonian euphaneropids (*Euphanerops*, *Cornovichthys*, *Achanarella*; Fig. 2n),

whose morphology is now best known from well-preserved 380-Myr-old *Euphanerops* material from the Late Devonian Miguasha Lagerstätte in Canada. Young individuals of *Euphanerops* are preserved as soft-tissue stains, but large individuals also show peculiar spongy calcifications of various elements of the endoskeleton (the internal, cartilaginous or bony skeleton of vertebrates), notably the fin radials, gill bars, vertebral elements, and elements that resemble the 'tongue' and annular cartilages of lampreys[61]. The most peculiar feature of *Euphanerops* is the large number (about 30 pairs) of gill bars that form its lamprey-like gill basket and extend back to the anal region. This is confirmed by a three dimensionally preserved specimen that shows impressions of the gill filaments[62]. Besides this feature, the overall appearance of *Euphanerops* resembles that of an anaspid, a group of Silurian–Devonian 'ostracoderms' that were long thought to be ancestral to lampreys, but are now regarded as being among the basal-most stem gnathostomes[60,61,63] (Fig. 4). Like anaspids, *Euphanerops* displays a long, posteroventrally slanting tail and a large anal fin, suggested to be paired — a unique case among vertebrates[64]. However, this requires confirmation, as does the elongate, paired ventrolateral fins that seem to have extended ventrally to the gill basket[61]. Whatever their relationships to *Jamoytius*, euphaneropids did not possess mineralized scales, but do have some endoskeletal characters uniquely shared with lampreys[65].

Finally, *Palaeospondylus*, from the Middle Devonian (390 Ma) of Scotland is still the most enigmatic early vertebrate, although it is known by hundreds of specimens. It is not preserved as a mere imprint, but clearly displays a vertebral column, a caudal fin with radials and fin supports, possible paired appendages, and its skull consists of several peculiar skeletal elements that cannot be clearly homologized with classic components of the vertebrate skull, be it a cyclostome or a gnathostome[66,67]. All of its skeletal elements are exclusively made up of a spongy calcified matter, which resembles that of the calcified endoskeleton of *Euphanerops*[61], and

---

BOX 1

Fossils and 'ancestors'

When the first description of the myllokunmingiids was published[39], early vertebrate palaeontologists were struck by the resemblance between these Lower Cambrian soft-bodied fossils from Chengjiang and various imaginary reconstructions of an ancestral vertebrate published during the twentieth century. For example, myllokunmingiids surprisingly resemble this imaginary reconstruction of an 'ancestral cephalochordate' (amphioxus) (see Figure) published at a time when some zoologists considered the absence of a complex head in living cephalochordates could be secondary. This reconstruction is a curious mix of a rather vertebrate-like, and even a 'ostracoderm'-like head, and some cephalochordate characters. It was thus intended to suggest that the overall morphology of the common ancestor to cephalochordates and vertebrates was rather vertebrate-like. Do such reconstructions of an entirely hypothetical 'ancestor', essentially based on inferences from extant and some fossil vertebrates, influence the way we interpret odd and poorly preserved soft-bodied fossils? Or do such fossils lead us to search for such old and supposedly prophetic reconstructions to justify intuitions? Although palaeontologists try to take a cold look at characters, it is probable that such reconstructions, based on the tree of life in vogue at a given time, unconsciously affect the way researchers look at certain fossils and favour wishful thinking when in search of ancestors. This was probably also the case for the interpretation of *Pikaia*. Image adapted with permission from ref. 42.

is therefore interpreted as calcified cartilage. Its resemblance to embryonic cartilage of extant osteichthyans (bony jawed vertebrates) has even led to the suggestion that *Palaeospondylus* might be a peculiar bony fish that failed to develop bone[68]. The anatomy of *Palaeospondylus* has been described, and this 'fish' has been tentatively referred to as practically all major fossil and extant vertebrate groups: hagfishes, lampreys, 'placoderms' (extinct armoured jawed fish), chondrichthyans (sharks, rays and chimaeras), teleosts, lungfish larvae and amphibian tadpoles[67,68]. All these interpretations are either dismissed or still debated. However, data on hagfish skeletal development[69] seem to enhance the superficial resemblance, already alluded to by some early authors, between the arrangement of certain elements of the *Palaeospondylus* skull and that of the cranial cartilages of late hagfish embryos. In addition, developmental data suggest that the absence of vertebral elements in hagfishes is probably secondary[70], and the vertebral column of *Palaeospondylus* may thus not preclude close relationships to hagfishes. Yet, no unambiguous character seems to be uniquely shared by hagfishes and *Palaeospondylus*.

These presumed soft-bodied chordates and vertebrates that were mostly devoid of hard tissue, except for occasional calcified cartilage, are generally collapsed and preserved as traces of variously transformed soft tissue[3,6]. Their reconstruction in 3D is often difficult, even by means of sophisticated techniques[60], and their descriptions are characteristically cautious. Should we simply forget about them? Do they provide us with any useful information? Or are they merely material support to our imagination, which is in turn guided by current views about the interrelationships of living animal groups? The art of reconstruction for palaeontologists is usually to put flesh on bones, but it is difficult when there is only decayed flesh and no bone! However, it is worth trying.

Hard-tissue data

Early vertebrate hard tissues are reputedly easier to identify. Their structure can be studied in detail by means of material or virtual (microtomographic) sections, eventually in 3D[71], and classic scanning electron microscopy techniques. Their characteristics can then be compared with those of living or more recent and well-known species. Nevertheless, palaeontologists are confronted with many of the same problems as for soft-tissue preservations when dealing with the earliest presumed vertebrate skeletal remains. The first clues to vertebrate hard tissues are that they are made of bioapatite; the tissues often show an ornamentation of tubercles (odontodes), or ridges, with a structure that resembles that of our teeth;

they have dentine that contains thin canals for cell processes; eventually enamel (enameloid) is present; and there is a pulp cavity (Fig. 3a). Other useful characters may be the surface ultra-sculpture, the small spaces that housed bone cells, and the grooves or canals that housed lateral-line sense organs. The exoskeleton of the earliest, articulated and duly recognized vertebrates, such as arandaspids or astraspids (Fig. 3b, c), show at least some of these characters[13,72,73]. However, younger vertebrates known from complete specimens, such as the Silurian and Devonian anaspids or galeaspids[73,74], lack dentine, and many of the Cambrian to Silurian 'microremains', referred to as vertebrates owing to the aspect of their ornamentation or their scale-like shape, lack some of these characteristic tissues. Instead, they show other hard tissues that no longer exist, such as lamelline (acellular dentine)[8,73]. Therefore, the earliest evidence for possible vertebrate hard-tissue remains are barely less puzzling than the Cambrian soft-bodied animals.

The first controversy about these problematic skeletal fragments arose with the discovery of *Anatolepis* from the Lower Ordovician and Upper Cambrian[75,76]. *Anatolepis* is represented by minute phosphatic fragments ornamented with elongate tubercles (Fig. 3d), which vaguely resemble the exoskeletal ornamentation of certain Silurian–Devonian ostracoderms, notably heterostracans (Fig. 4). Therefore, *Anatolepis* was first regarded as a possible heterostracan; this was immediately contested by some, whereas others considered it plausible. At around the same time, *Anatolepis* was tentatively referred to an arthropod, but again this raised debate. Later studies of the tissue structure of these fragments using new techniques showed that the tubercles of *Anatolepis* were in fact hollowed by a pulp cavity capped by a somewhat dentine-like tissue, and connected by a lamellar tissue, which was perforated by thin vertical canals (Fig. 3e)[77]. Nevertheless, these new data failed to convince the sceptics[78]. *Anatolepis* may remain an enigma — as long as no articulated individual turns up. Other alleged Late Cambrian vertebrate bone fragments have been described from Australia[79] and superficially resemble the exoskeletal bone ornamentation of the Ordovician arandaspid *Porophoraspis*[11]; however, they are also strikingly similar to some Palaeozoic arthropod carapaces[78]. In sum, apart from the euconodonts, whose possible vertebrate affinities essentially rest on soft-tissue characters, there is no undisputed evidence for Cambrian vertebrates that possess a mineralized skeleton. By contrast, the following Ordovician period not only yields articulated vertebrates covered with extensive mineralized armour and scales, but also numerous isolated bone fragments and scales[80]. Most of these microremains, such as
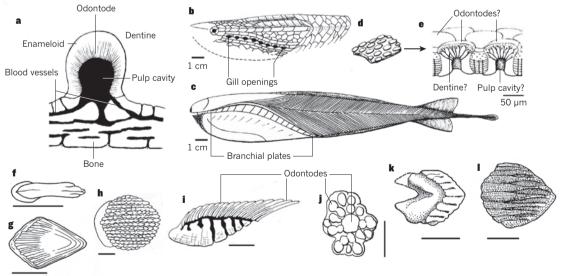
Figure 3 | Late Cambrian, Ordovician and early Silurian vertebrate exoskeletons. **a**, Ideal vertical section through a typical, ornamented vertebrate exoskeleton showing a tubercle (odontode) attached to a bony base (not to scale). **b, c**, The most complete articulated Ordovician vertebrates, *Astraspis* (**b**) and *Sacabambaspis* (**c**) (adapted from refs 86, 93). **d, e**, Exoskeleton fragment of the debated vertebrate *Anatolepis* (**d**) and vertical section of the possible odontodes (**e**) (adapted from ref. 77). **f–l**, Major types of isolated vertebrate scales retrieved from Upper Ordovician and Lower Silurian rocks: a thelodont (**f**), an 'acanthodian' (**g**), the possible chondrichthyan *Mongolepis* as an external view (**h**) and vertical section (**i**), and the vertebrates of uncertain affinities *Tesakoviaspis* (**j**), *Apedolepis* (**k**) and *Areyongalepis* (**l**). (**f–j** adapted from ref. 8 and **k, l** from ref. 82) Scale bars are 1 cm (**b, c**), 0.5 mm (**f–l**) and 50 μm (**e**).
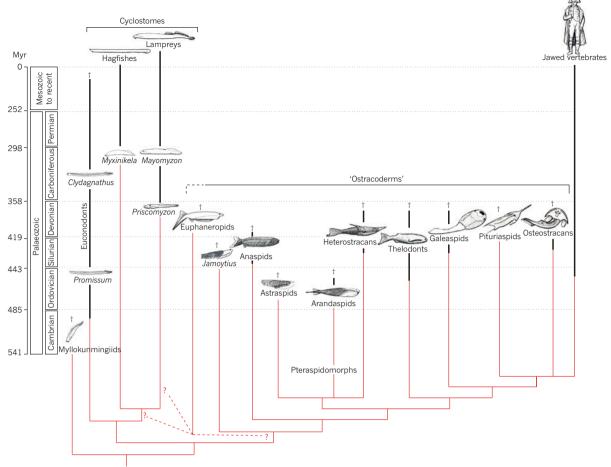
Figure 4 | Distribution through geological time (black bars), and patterns of interrelationships (red) of the major Palaeozoic jawless vertebrate groups and their extant relatives. Pattern of relationships adapted from ref. 60, except for the position of the euphaneropids. *Promissum* and *Clydagnathus* adapted from ref. 47; lampreys, hagfishes, *Myxinikela*, *Mayomyzon*, *Priscomyzon*, Euphaneropids, Anaspids, Heterostracans and Thelodonts adapted from ref. 65; *Jamoytius* adapted from ref. 60; *Astraspids* adapted from ref. 86; Arandaspids adapted from ref. 93. Not to scale. †Extinct groups.

Skiichthys[81] (a possible 'placoderm') or other scale-like elements, show at least some hard-tissue characters that are shared with younger vertebrate groups. However, others, such as *Areyongalepis* (Fig. 3k) and *Apedolepis* (Fig. 3l) are very puzzling[82]. Isolated vertebrate remains occur sporadically throughout most of the Ordovician and early Silurian and, despite their amazing diversity of hard-tissue structures, show an increasingly close resemblance to structures and ornamentations of the late Silurian and Devonian vertebrate groups, which are known from complete skeletons.

The three articulated Ordovician vertebrates, *Astraspis* (Fig. 3b), *Arandaspis* and *Sacabambaspis* (Fig. 3c)[11-13], and the bone assemblages of *Eriptychius*[83] and *Ritchieichthys*[84] show the overall morphology of the earliest vertebrates that have an extensive exoskeleton with a large head shield composed of either large plates or polygonal platelets, a posteriorly slanting series of numerous gill openings, and a scale-covered body and tail[85,86]. However, they provide no information about internal anatomy, apart from uninformative fragments of calcified cartilage in *Eriptychius*[83], and faint internal impressions of the gill pouches in *Astraspis* and *Sacabambaspis*. Orbits indicate the presence of eyes, and paired dorsal openings in arandaspids are interpreted as pineal foramina, but the position of nasal openings is unclear[57]. The lower lip of arandaspids is covered with a series of minute platelets, suggesting a filtering function, as in the younger heterostracans[87]. These articulated fossils may give the impression that all Ordovician fishes looked like big armoured tadpoles. However, the diversity of the scales and other microremains retrieved from coeval Ordovician rocks suggests that different morphologies may have existed already. *Porophoraspis* is regarded as an arandaspid, but some relatively large plates referred as this genus are difficult to reconcile with the head-skeleton morphology of either *Arandaspis* or *Sacabambaspis*[11]. Among the isolated scales retrieved from Ordovician and Early Silurian rocks, some clearly belong to thelodonts (a group of 'ostracoderms'; Figs 3f, 4) and 'acanthodians' (Fig. 3g; presumed stem chondrichthyans). Both of these were known later by complete specimens, whereas others, such as *Mongolepis* (Fig. 3h, i), *Teslepis*, *Sodolepis* and *Tesakoviaspis* (Fig. 3j)[8], all presumed chondrichthyans (shark relatives), and still-unnamed forms[80] may have belonged to vertebrates that had an entirely micromeric (composed of minute scales) exoskeleton like that of sharks. Their body structure will remain unknown unless articulated material is discovered in some still-elusive Lagerstätte. Although some of these scales are, by default, referred to as chondrichthyans, they are in fact vertebrates in limbo.

After the Middle Ordovician, no articulated vertebrate turns up until the mid-Silurian (around 433 Ma), apart from the Late Ordovician euconodont *Promissum*[45]. Then, relatively complete representatives of the six major Silurian–Devonian 'ostracoderm' groups (anaspids, heterostracans, thelodonts, galeaspids, pituriaspids and osteostracans; Fig. 4) occur, and, shortly after (about 430 Ma) the earliest complete jawed vertebrates, notably 'placoderms'[10], 'acanthodians' and osteichthyans (bony fishes)[9]. Such articulated or well-preserved material is generally the key to suggesting a systematic position for some of the microremains from the Ordovician and early Silurian, and tracing back the distribution of these major groups through time (Fig. 4). Moreover, the number of anatomical characters that this material now offers us allows for better supported reconstructions of the interrelationships of these groups.

The phylogenetic trees of fossil and living vertebrates generally agree on the position of the 'ostracoderms' as a series of jawless stem gnathostomes,

with galeaspids, osteostracans (and possibly pituriaspids) as successive sister groups of the jawed vertebrates[50,57,60,88] (Fig. 4). This is partly because galeaspids and osteostracans have an extensively calcified or ossified endoskeleton, which preserves the cavities and canals that housed the brain, sensory capsules, nerves and blood vessels, including the pectoral girdles and fins in osteostracans, thereby providing a wealth of anatomical characters that can be compared with their homologues in jawed vertebrates[57,88,89]. However, the relationships of other 'ostracoderm' groups is poorly supported because they are devoid of a calcified endoskeleton, and their exoskeleton, which is sometimes entirely micromeric, provides indirect information about their internal anatomy in the form of faint impressions of, for example, gill pouches, brain, olfactory organs or labyrinth[57]. As is the case for heterostracans, but there are no data for anaspids, and only a few thelodonts provide some information[57,90,91]. Heterostracans are characterized by a single pair of common branchial openings, and are gathered with astraspids and arandaspids in the pteraspidomorphs (Fig. 4)[57,72]. However, apart from the presence of large median dorsal and ventral head plates made of acellular bone, and a similar honeycomb-like layer in the exoskeleton of heterostracans and arandaspids, shared derived characters that are unique to these three groups are scarce.

For almost a century, most debates about the relationships of the various 'ostracoderm' groups have been centred on the structure of the rostral part of the head: the olfactory organs, their relation to the hypophysis (pituitary) and the oral region. Classically, the dorsal position of the common nasal and hypophyseal duct of osteostracans and anaspids was compared with the condition in lampreys[92,93]. However, the recent description of the same region of the head in galeaspids has provided new insights[94]. The still elusive heterostracan and thelodont internal anatomy could possibly be reconstructed on the basis of that of galeaspids, with paired nasal sacs and an anteriorly directed hypophyseal duct. This would mean that a galeaspid-like anatomy might have been widespread among stem gnathostomes, and that the allegedly lamprey-like nasohypophyseal complex of osteostracans is independently derived from such a condition.

Fossils, phylogeny and technologies

It is sometimes said that fossils never, or rarely, overturn patterns of relationships based on extant organisms. Patterson[95] mentioned a few possible exceptions, notably the 'calcichordate theory'[96], which assumed that an ensemble of Palaeozoic echinoderm-like groups classically referred to as stylophorans are a paraphyletic array of stem chordates, stem cephalochordates, stem tunicates and stem vertebrates, the calcitic skeleton of which has been lost several times. It also suggested that tunicates, and not cephalochordates, were the closest extant relatives of vertebrates (contra to the then accepted relationships). This theory has raised heated controversies[97], but all stylophorans are now regarded as stem echinoderms. However, recent molecular phylogenies strongly support this tunicate–vertebrate relationship[98]. Tunicates and vertebrates are therefore gathered in a group called Olfactores, a name that, paradoxically, was erected in the framework of the calcichordate theory[96], because some stylophorans that were thought to be stem tunicates display internal structures that resemble vertebrate olfactory organs. Patterson[95] predicted that molecular sequence data would be the best test of the 'calcichordate theory', and, coincidentally, the test seems to have been positive regarding tunicate relationships.

Regarding vertebrates, the hypothesis of living cyclostome paraphyly (that lampreys are more closely related to gnathostomes than to hagfishes) was only based on phenotypic data derived from extant species[99]. Palaeontological data have been merely adapted to this pattern of relationships, because of the long-lasting conviction that certain 'ostracoderms' (osteostracans and anaspids) were most closely related to lampreys[57,93]. More accurate character analyses later showed that 'ostracoderms' were exclusively stem gnathostomes, and the recent revival of cyclostome monophyly had no major bearing on their interrelationships[60]. None of the fossils discussed earlier, be they soft-body imprints, bone fragments, scales or articulated skeletons, seems currently liable to overturn the interrelationships of the major extant vertebrate groups. However, they provide a minimal age for certain characters (thus the groups they define), and

may reveal unsuspected character combinations that allow the reconstruction of the stepwise assembly of novel body plans that foreshadow major evolutionary transitions. This is, for example, what 'ostracoderms' document with the succession of characters that make up the jawed vertebrate body: the rise of the exoskeleton, cellular bone, endoskeletal bone, enlarged cerebellum or pectoral fins, but they are still rather powerless in providing a scenario for the rise of jaws, which is largely left in the hands of evolutionary developmental biologists. Nevertheless, the recent consideration of braincase anatomy in the basal-most 'placoderms' suggests that the anatomical gap between such 'ostracoderms' as galeaspids and osteostracans, and the earliest jawed vertebrates, may not have been that large, and that the prerequisites to the rise of jaws were already there[100].

The future of early vertebrate palaeontology rests on the quality of the data it can provide, especially on fossils derived from crucial periods, such as the Late Cambrian, Early Ordovician and early Silurian. Early vertebrates are generally difficult material, compressed or crushed in hard rocks. Throughout the twentieth century, some early vertebrate palaeontologists gave much weight to the then new preparation techniques[57]. Nowadays, they would be amazed by the quality of the data obtained from high-resolution X-ray microtomography. Also, soft tissues preserved as mere stains can be studied by element mapping that provides information on the fossilization process and sometimes the nature of the preserved tissues themselves. Armed with these non-destructive techniques, early vertebrate palaeontologists can considerably refine their observations and must not be afraid of proposing audacious interpretations of these miserable remains, even though 'squashed slugs' may be slippery! ∎

1. Hou, X.-G. et al. The Cambrian Fossils of Chengjiang, China: the Flowering of Early Animal Life (Blackwell Publishing, 2007).
2. Royal Ontario Museum. The Burgess Shale http://www.burgess-shale.rom.on.ca (Royal Ontario Museum, 2011).
3. Briggs, D. E. G. The role of decay and mineralization in the preservation of soft-bodied fossils. Annu. Rev. Earth Planet. Sci. 31, 275–301 (2003).
4. Sansom, R. S., Gabbott, S. E. & Purnell, M. A. Decay of chordate characters causes bias in fossil interpretation. Nature 463, 797–800 (2010).
5. Sansom, R. S. Gabbott, S. E. & Purnell M. A. Decay of vertebrate characters in hagfish and lamprey (Cyclostomata) and the implications for the vertebrate fossil record. Proc. R. Soc. Lond. B 278, 1150–1157 (2011).
6. Sansom, R. S., Gabbott, S. E. & Purnell, M. A. Atlas of vertebrate decay: a visual and taphonomic guide to fossil interpretation. Palaeontology 56, 457–474 (2013).
7. Turner, S. in Advances in the Origin and Early Radiation of Vertebrates (eds Arratia, G., Wilson, M.V.H. & Cloutier, R.) 67–94 (Pfeil, 2004).
8. Karatayute-Talimaa, V. N. Determination methods for the exoskeletal remains of early vertebrates. Fossil Record 1, 21–51 (1998).
   An important review of the tissue structure in Early Palaeozoic microremains.
9. Zhu, M., Zhao, W., Jia, L., Qiao, T. & Qu, Q. The oldest articulated osteichthyan reveals mosaic gnathostome characters. Nature 458, 469–474 (2009).
10. Zhu, M. et al. A Silurian placoderm with osteichthyan-like marginal jaw bones. Nature 502, 188–193 (2013).
11. Ritchie, A. & Gilbert-Tomlinson, J. First Ordovician vertebrates from the southern hemisphere. Alcheringa 1, 351–368 (1977).
12. Gagnier, P.-Y. Sacabambaspis janvieri, vertébré ordovicien de Bolivie. 1, analyse morphologique [in French]. Ann. Paleontol. 79, 19–69 (1993).
13. Sansom, I. J., Smith, M. P., Smith, M. M. & Turner, P. Astraspis — the anatomy and histology of an Ordovician fish. Palaeontology 40, 625–643 (1997).
14. Donoghue, P. C. J. & Purnell, M. A. Genome duplication, extinction and vertebrate evolution. Trends Ecol. Evol. 20, 312–319 (2005).
15. Bardack, D. & Zangerl, R. First fossil lamprey: a record from the Pennsylvanian of Illinois. Science 162, 1265–1267 (1968).
16. Janvier, P. & Lund, R. Hardistiella montanensis N. gen. et sp. (Petromyzontida) from the Lower Carboniferous of Montana, with remarks on the affinities of lampreys. J. Vert. Paleontol. 2, 407–413 (1983).
17. Gess, R. W., Coates, M. I. & Rubidge, B. S. A lamprey from the Devonian of South Africa. Nature 443, 981–984 (2006).
18. Chang, M. M., Zhang, J. & Miao, D. A lamprey from the Cretaceous Jehol biota of China. Nature 441, 972–974 (2006).
19. Bardack, D. First fossil hagfish (Myxinoidea): a record from the Pennsylvanian of Illinois. Science 254, 701–703 (1991).
20. Poplin, C., Sotty, D. & Janvier, P. Un Myxinoïde (Craniata, Hyperotreti) dans le Konservat-Lagerstätte Carbonifère supérieur de Montceau-les-Mines (Allier, France) [in French]. CR. Acad. Sci. II A 332, 345–350 (2001).
21. Graham, A., Butts, T., Lumsden, A. & Kiecker, C. What can vertebrates tell us about segmentation? EvoDevo 5, 24 (2014).
22. Conway Morris, S. & Caron, J.-B. Pikaia gracilens Walcott, a stem-group chordate from the Middle Cambrian of British Columbia. Biol. Rev. Camb. Philos. Soc. 87, 480–512 (2012).
23. Gould, S. J. Wonderful Life: The Burgess Shale and the Nature of History (W.W. Norton & Co, 1990).

24. Chen, J. Y., Dzik, J., Edgecombe, G. D., Ramskjöld, L. & Zhou, G. Q. A possible Early Cambrian chordate. *Nature* **377,** 720–722 (2002).
25. Shu, D. et al. A new species of Yunnanozoan with implications for deuterostome evolution. *Science* **299,** 1380–1384 (2003).
26. Chen, J.-Y., Huang, D.-Y. & Li, C.-W. An early Cambrian craniate-like chordate. *Nature* **402,** 518–522 (1999).
27. Holland, N. D. & Chen, J. Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. *Bioessays* **23,** 142–151 (2001).
28. Mallatt, J. & Chen, J. Fossil sister group of craniates: predicted and found. *J. Morphol.* **258,** 1–31 (2003).
   The most detailed description of Yunnanozoan anatomy.
29. Mallatt, J., Chen, J. & Holland, N. D. Comment on "A new species of ynnanozoan, with implications for deuterostome evolution". *Science* **300,** 1372 (2003).
30. Shu, D. & Conway Morris, S. Response to Comment on "A new species of yunna-nozoan with implications for deuterostome evolution". *Science* **300,** 1372 (2003).
31. Shu, D.-G. et al. Primitive deuterostomes from the Chengjiang Lagerstätte (Lower Cambrian, China). *Nature* **414,** 419–424 (2001).
32. Gee, H. On being vetulicolian. *Nature* **414,** 407–409 (2001).
33. Caron, J.-B. *Banffia constricta,* a putative vetulicolid from the Middle Cambrian Burgess Shale. *Trans. R. Soc. Edinb. Earth Sci.* **96,** 95–111 (2005).
34. Shu, D.-G., Conway Morris, S. & Zhang, X.-L. A *Pikaia*-like chordate from the Lower Cambrian of China. *Nature* **384,** 157–158 (1996).
35. Shu, D.-G., Chen, L., Han, J. & Zhang, X. L. An Early Cambian tunicate from China. *Nature* **411,** 472–473 (2001).
36. Chen, J. Y. et al. The first tunicate from the early Cambrian of South China. *Proc. Natl Acad. Sci. USA* **100,** 8314–8318 (2003).
37. Ritchie, A. *Ainiktozoon loganense* Scourfield, a protochordate? From the Silurian of Scotland. *Alcheringa* **9,** 117–142 (1985).
38. Van der Brugghen, W., Schram, F. R. & Martill, D. M. The fossil *Ainiktozoon* is an Arthropod. *Nature* **385,** 589–590 (1997).
39. Shu, D.-G. et al. Lower Cambrian vertebrates from South China. *Nature* **402,** 42–46 (1999).
40. Shu, D. G. et al. Head and backbone of the Early Cambrian vertebrate *Haikouichthys. Nature* **421,** 526–529 (2003).
41. Conway Morris, S. & Caron, J.-B. A primitive fish from the Cambrian of North America. *Nature* **512,** 419–422 (2014).
42. Holmgren, N. & Stensiö, E. in *Handbuch des Vergleichenden Anatomie der Wirbeltiere* [in German] Vol. 4 (eds Bolk, L., Göppert, E., Kallius, E. & Lubosch, W.) 233–500 (Urban & Schwarzenberg, 1936).
43. Murdock, D. J. E. et al. The origin of conodonts and of vertebrate mineralized skeletons. *Nature* **502,** 546–549 (2013).
44. Briggs, D. E. G., Clarkson, E. N. K. & Aldridge, R. J. The conodont animal. *Lethaia* **20,** 1–14 (1983).
45. Gabbott, S. E., Aldridge, R. J. & Theron, J. N. A giant conodont with preserved muscle tissue from the Upper Ordovician of South Africa. *Nature* **374,** 800–803 (1994).
46. Aldridge, R. J., Briggs, D. E. G., Smith, M. P., Clarkson, E. N. K. & Clark, D. N. L. The anatomy of conodonts. *Philos. Trans. R. Soc. Lond. B* **340,** 405–421 (1993).
47. Purnell, M. A. Large eyes and vision in conodonts. *Lethaia* **28,** 187–188 (1995).
48. Donoghue, P. C. J. Growth and patterning in the conodont skeleton. *Philos. Trans. R. Soc. Lond. B* **353,** 633–666 (1998).
49. Turner, S. et al. False teeth: conodont-vertebrate phylogenetic relationships revisited. *Geodiversitas* **32,** 545–594 (2010).
50. Donoghue, P. C. J., Forey, P. L. & Aldridge, R. J. Conodont affinity and chordate phylogeny. *Biol. Rev. Camb. Philos. Soc.* **75,** 191–251 (2000).
51. Kreijsa, R. J., Bringas, P. & Slavkin, H. A neontological interpretation of conodont elements based on agnathan cyclostome tooth structure, function and development. *Lethaia* **23,** 359–378 (1990).
52. Dzik, J. Conodont affinity of the enigmatic Carboniferous chordate *Conopiscius. Lethaia* **42,** 31–38 (2009).
53. Goudemand, N., Orchard, M. J., Urdy, S., Bucher, H. & Tafforeau, P. Synchrotron-aided reconstruction of the conodont feeding apparatus and implications for the mouth of the first vertebrates. *Proc. Natl Acad. Sci. USA* **108,** 8720–8724 (2011).
54. Briggs, D. E. G. & Clarkson, E. N. K. An enigmatic chordate from the Lower Carboniferous Granton "shrimp-bed" of the Edinburgh district, Scotland. *Lethaia* **20,** 107–115 (1987).
55. Aldridge, R. J. & Donoghue, P. C. J. in *The Biology of Hagfishes* (eds Jørgensen, J. M., Lomholt, J. P., Weber, R. E. & Malte, H.) 16–31 (Chapman and Hall, 1998).
56. Bardack, D. & Richardson, E. S. Jr. New agnathous fishes from the Pennsylvanian of Illinois. *Fieldiana: Geology* **33,** 489–510 (1977).
57. Janvier, P. *Early Vertebrates* (Oxford Univ. Press, 1996).
   This text is a general overview of early vertebrate anatomy and relationships.
58. Bardack, D. in *The Biology of Hagfishes* (eds Jørgensen, J. M., Lomholt, J. P., Weber, R. E. & Malte, H.) 3–14 (Chapman and Hall, London, 1998).
59. White, E. I. *Jamoytius kerwoodi*, a new chordate from the Silurian of Lanarkshire. *Geol. Mag.* **83,** 89–97 (1946).
60. Sansom, R. S., Freedman, K., Gabbott, S. E., Aldridge, R. J. & Purnell, M. A. Taphonomy and affinity of an enigmatic Silurian vertebrate, *Jamoytius kerwoodi* White. *Palaeontology* **53,** 1393–1409 (2010).
61. Janvier, P. & Arsenault, M. The anatomy of *Euphanerops longaevus* Woodward, 1900, an anaspid-like jawless vertebrate from the Upper Devonian of Miguasha, Quebec, Canada. *Geodiversitas* **29,** 143–216 (2007).
62. Janvier, P., Desbiens, S. & Willett, J. A. & Arsenault, M. Lamprey-like gills in a gnatho-stome-related Devonian jawless vertebrate. *Nature* **440,** 1183–1185 (2006).
63. Blom, H. New birkeniid anaspid from the Lower Devonian of Scotland and its phylogenetic implications. *Palaeontology* **55,** 641–652 (2012).
64. Sansom, R. S., Gabbott, S. E. & Purnell, M. A. Unusual anal fin in a Devonian jawless vertebrate reveals complex origins of paired appendages. *Biol. Lett.* **9,** 20130002 (2013).
65. Janvier, P. Modern look for ancient lamprey. *Nature* **443,** 921–924 (2006).
66. Sollas, W. J. & Sollas, I. B. An account of the Devonian fish, *Palaeospondylus gunni,* Traquair. *Phil. Trans. R. Soc. Lond. B* **196,** 267–294 (1904).
67. Moy Thomas, J. A. The Devonian fish *Palaeospondylus gunni* Traquair. *Philos. Trans. R. Soc. Lond. B* **230,** 391–413 (1940).
68. Johanson, Z., Kearsley, A., Den Blaauwen, J., Newman, M. & Smith, M. M. No bone about it: an enigmatic Devonian fossil reveals a new skeletal framework – a potential loss of gene regulation. *Semin. Cell Dev. Biol.* **21,** 414–423 (2010).
69. Oisi, Y., Ota, K. G., Fujimoto, S. & Kuratani, S. S. Development of the chondrocranium in hagfishes, with special reference to the early evolution of vertebrates. *Zoolog. Sci.* **30,** 944–961 (2013).
70. Janvier, P. All vertebrates do have vertebrae. *Curr. Biol.* **21,** R661–R663 (2011).
71. Sanchez, S., Ahlberg, P. E., Trinajstic, K. M., Mirone, A. & Tafforeau, P. Three-dimensional synchrotron virtual palaeohistory: A new insight into the world of fossil bone microstructure. *Microsc. Microanal.* **18,** 1095–1105 (2012).
72. Sansom, I. J., Donoghue, P. C. J. & Albanesi, G. Histology and affinity of the earliest armoured vertebrate. *Biol. Lett.* **1,** 446–449 (2005).
73. Sire, J.-Y., Donoghue, P. C. J. & Vikaryous, M. K. Origin and evolution of the integumentary skeleton in non-tetrapod vertebrates. *J. Anat.* **214,** 409–440 (2009).
   This is an updated review of exoskeletal hard tissues in early fishes.
74. Wang, N.-Z., Donoghue, P. C. J., Smith, M. M. & Sansom, I. J. Histology of the galeaspid dermoskeleton and endoskeleton, and the origin and early evolution of the vertebrate cranial endoskeleton. *J. Vert. Paleontol.* **25,** 745–756 (2005).
75. Bockelie, T. G. & Fortey, R. A. An early Ordovician vertebrate. *Nature* **260,** 36–38 (1976).
76. Repetski, J. E. A fish from the Upper Cambrian of North America. *Science* **200,** 529–531 (1978).
77. Smith, M. P., Sansom, I. J. & Repetski, J. E. Histology of the first fish. *Nature* **380,** 702–704 (1996).
78. Smith, M. P., Sansom, I. J. & Cochrane, K. D. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 67–84 (Taylor and Francis, 2001).
79. Young, G. C., Karatayute-Talimaa, V. N. & Smith, M. M. A possible Late Cambrian vertebrate from Australia. *Nature* **383,** 810–812 (1996).
80. Sansom, I. J., Smith, M. M. & Smith, M. P. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 156–171 (Taylor and Francis, 2001).
81. Smith, M. M. & Sansom, I. J. Exoskeletal micro-remains of an Ordovician fish from the Harding Sandstone of Colorado. *Palaeontology* **40,** 645–658 (1997).
82. Young, G. C. Ordovician microvertebrate remains from the Amadeus Basin, central Australia. *J. Vert. Paleontol.* **17,** 1–25 (1997).
83. Denison, R. H. Ordovician vertebrates from Western United States. *Fieldiana: Geology* **16,** 131–192 (1967).
84. Sansom, I. J., Haines, P. W., Andreev, P. & Nicoll, R. S. A new pteraspidomorph from the Nibil Formation (Katian, Late Ordovician) of the Canning Basin, Western Australia. *J. Vert. Paleontol.* **33,** 764–769 (2013).
85. Pradel, A., Sansom, I. J., Gagnier, P.-Y., Cespedes, R. & Janvier, P. The tail of the Ordovician fish Sacabambaspis. *Biol. Lett.* **3,** 72–75 (2007).
86. Janvier, P. in *Recent Advances in the Origin and Early Radiation of Vertebrates* (eds Arratia, G., Wilson, M. V. H. & Cloutier, R.) 29–52 (Pfeil, 2004).
87. Purnell, M. A. Feeding in extinct jawless heterostracan fishes and testing scenarios of early vertebrate evolution. *Proc. R. Soc.* **269,** 83–88 (2002).
88. Janvier, P. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 172–186 (Taylor and Francis, 2001).
89. Janvier, P., Arsenault, M. & Desbiens, S. Calcified cartilage in the paired fins of the osteostracan *Escuminaspis laticeps* (Traquair 1880), from the Late Devonian of Miguasha (Québec, Canada), with a consideration of the early evolution of the pectoral fin endoskeleton in vertebrates. *J. Vert. Paleontol.* **24,** 773–779 (2004).
90. Donoghue, P. C. J. & Smith, M. P. The anatomy of *Turinia pagei* (Powrie) and the phylogenetic status of the Thelodonti. *Trans. R. Soc. Edinb. Earth Sci.* **92,** 15–37 (2001).
91. Märss, T. & Wilson, M. V. H. Buccopharyngo-branchial denticles of *Phlebolepis elegans* Pander (Thelodonti, Agnatha). *J. Vert. Paleontol.* **23,** 601–612 (2008).
92. Stensiö, E. A. The Devonian and Downtonian vertebrates of Spitsbergen. 1. Family Cephalaspidae. *Skr. Svalbard Ishav.* **12,** 1–391 (1927).
93. Janvier, P. Early jawless vertebrates and cyclostome origins. *Zoolog. Sci.* **25,** 1045–1056 (2008).
94. Gai, Z., Donoghue, P. C. J., Zhu, M., Janvier, P. & Stampanoni, M. Fossil jawless fish from China foreshadows early jawed vertebrate anatomy. *Nature* **476,** 324–327 (2011).
95. Patterson, C. Significance of fossils in determining evolutionary relationships. *Annu. Rev. Ecol. Syst.* **12,** 195–223 (1981).
96. Jefferies, R. P. S. *The Ancestry of the Vertebrates* (British Museum (Natural History), 1996).
   This is an extensive account of the calcichordate theory.
97. Gee, H. *Before the Backbone: Views on the Origin of the Vertebrates* (Chapman & Hall, 1996).
   This text is a review of fossil-based theories about the origin of vertebrates.
98. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicate and not cephalochordate are the closest living relatives of vertebrates. *Nature* **439,** 965–968 (2006).
99. Løvtrup, S. *The Phylogeny of Vertebrata* (Wiley, 1977).
100. Dupret, V., Sanchez, S., Goujet, D., Tafforeau, P. & Ahlberg, P. E. A primitive placoderm sheds light on the origin of the jawed vertebrate face. *Nature* **507,** 500–503 (2014).

REVIEW

The origin and early phylogenetic history of jawed vertebrates

Martin D. Brazeau[1] & Matt Friedman[2]

Fossils of early gnathostomes (or jawed vertebrates) have been the focus of study for nearly two centuries. They yield key clues about the evolutionary assembly of the group's common body plan, as well the divergence of the two living gnathostome lineages: the cartilaginous and bony vertebrates. A series of remarkable new palaeontological discoveries, analytical advances and innovative reinterpretations of existing fossil archives have fundamentally altered a decades–old consensus on the relationships of extinct gnathostomes, delivering a new evolutionary framework for exploring major questions that remain unanswered, including the origin of jaws.

Jawed vertebrates (gnathostomes) comprise more than 99% of living vertebrate species, including humans. This diversity is built on features including jaws, teeth, paired appendages, and specialized embryonic and skeletal tissues (Box 1); centuries of research have attempted to explain their origins[1–7]. In particular, jaws and paired appendages have become flagship systems in the study of evolutionary novelty[5,7] — a key research programme in evolutionary biology[8].

The deepest split in the modern gnathostome tree is that between the chondrichthyans (sharks, rays and chimaeras) and the osteichthyans (bony fishes and tetrapods). This divergence occurred in the Palaeozoic era, at least 423 million years ago (Ma)[9], leaving a vast temporal and evolutionary gulf between modern lineages, with ample time for new innovations to overwrite primitive conditions. These complexities compel researchers to turn to the Palaeozoic fossil record to elucidate the origin of jawed vertebrates. A few well-preserved fossil taxa from a handful of Silurian–Permian sites in Europe and North America[10] shaped late nineteenth- and early twentieth-century hypotheses of gnathostome evolution[1,11,12] (Fig. 1). Many of these narratives persist to this day, either implicitly or explicitly. However, fossils once hailed as avatars for scenarios of jaw[12,13] or fin[1,14] origins often turn out to be specialized rather than primitive after phylogenetic investigation[15,16]. Until they are placed in a evolutionary tree, Palaeozoic fossils are mute on the question of gnathostome origins.

In this Review, we examine the progress made in the past two decades on the study of early gnathostome interrelationships, focusing on key fossil discoveries that have prompted a renewed intensity of phylogenetic investigation. Although tremendous advances have been made, much work remains before this research can deliver finely atomized transformational hypotheses such as those available for mammals[17], birds[18] and early tetrapods[19].

Phylogeny of extant gnathostomes

From the perspective of modern lineages alone, deep vertebrate phylogeny is well resolved and there is little disagreement about the branching patterns surrounding the gnathostome crown node (Box 1). Morphological[20] and molecular[21] data unambiguously indicate that chondrichthyans and osteichthyans are reciprocally monophyletic sister taxa. Together, they form a clade to the exclusion of the jawless cyclostomes: hagfishes and lampreys (Box 1). Molecular evidence strongly supports the monophyly of living agnathans with respect to jawed vertebrates. The

long-standing morphological hypothesis indicated the union of lampreys and gnathostomes to the exclusion of hagfishes[10,22], but re-appraisal of traits in living species[23–25] and reconsideration of existing data sets[26] have exposed its weaknesses.

These established relationships put the study of early gnathostome evolution at an advantage. Modern taxa can be organized into a set of crown groups delimiting three stem lineages: the respective branches subtending Osteichthyes and Chondrichthyes, and the branch subtending their last common ancestor (Box 1). The palaeontological problem is reduced to phylogenetic placement of Palaeozoic fossils within this three-branch framework.

Palaeozoic jawed vertebrates and their phylogeny

In this section we outline the range of early gnathostome diversity and review the recent history of progress on their phylogenetic relationships.

Diversity of Palaeozoic jawed vertebrates

Putative examples of jawed vertebrates date to the Ordovician period[27–29], but the first definitive remains are of early Silurian age[30]. Early Devonian (419 Ma) mandibulate gnathostomes were already ecologically diverse[31] and, by the close of the Devonian (360 Ma), the first tetrapods and many of their adaptations for terrestriality had emerged[19].

Early jawed fishes are divided into four broad categories: ancient representatives of chondrichthyans and osteichthyans, along with two exclusively extinct assemblages: acanthodians and placoderms. The early chondrichthyan record is dominated by isolated denticles (scales), teeth and spines. The oldest records of scales attributed to chondrichthyans are from the earliest Silurian (around 443 Ma)[27], such as mongolepids[32]. Sinacanthids, represented by isolated spines that share histological similarities with chondrichthyans[33], are also known from the early Silurian (about 438 Ma)[30]. The oldest universally accepted chondrichthyans are substantially younger, represented by Early Devonian body fossils (around 400 Ma; Fig. 2e). Some of these specimens derive from the 'Malvinokaffric Realm', a cold-water Southern Hemisphere palaeobiogeographic province that yields distinctive jawed vertebrate faunas almost exclusively composed of acanthodians and chondrichthyans[34]. Articulated chondrichthyans remain rare throughout the Devonian, with most specimens known from the exceptional Late Devonian Cleveland Shale Lagerstätte (Fig. 1).
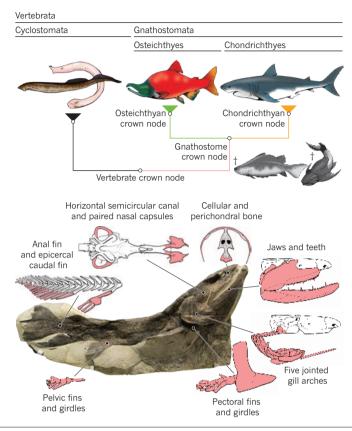
The late Silurian–Devonian osteichthyan record is considerably better than that of chondrichthyans owing to the armour of dermal plates and

[1]Department of Life Sciences, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, UK. [2]Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, UK.

BOX 1

Crowns, stems and the characters of jawed vertebrates

Crown-, total- and stem-group concepts provide a useful framework for navigating evolutionary trees that include fossils. The tree shown in the figure reflects the most basic splits among living vertebrates. Crown groups comprise the last common ancestor of a group of living species plus all of its descendants, both fossil and modern. The gnathostome crown group includes the last common ancestor of osteichthyans (represented by a salmon) and chondrichthyans (represented by a shark) plus all of its descendants, and comprises all the green and orange parts of the tree. Total groups include the crown group of interest plus all extinct forms more closely related to that lineage than to any other living species. Here, the gnathostome total group is represented by all coloured parts of the tree. Stem groups are equal to a clade's total group minus its crown group, shown here by the pink lineage connecting the vertebrate and gnathostome crown nodes. Jawed vertebrates include all of the gnathostome crown, and the upper reaches of the gnathostome stem. The lower part of the gnathostome stem is populated by jawless ostracoderms, which are more closely related to jawed vertebrates than they are to modern jawless fishes. The principal task faced by palaeontologists is to fit fossil groups (such as acanthodians and placoderms; the dagger symbol indicates that they are extinct) within the genealogical framework for modern species. Monophyly of jawed vertebrates is evidenced by a series of shared morphological specializations including, but not limited to, jaws. Key gnathostome features are illustrated here for *Eusthenopteron* (Cleveland Museum of Natural History CMNH 8158, image courtesy of D. Chapman), an osteichthyan and relative of land vertebrates. These traits must have evolved along the gnathostome stem lineage, but without fossils it is impossible to determine the order in which — or when — they arose.

ossified endoskeleton typical of bony fishes. Consequently, osteichthyans have been intensively studied, with particular emphasis on sarcopterygians (lobe-finned fishes), reflecting their importance in reconstructing early stages of tetrapod evolution[19,35,36]. Lobe fins are known from the late Silurian (about 423 Ma)[9], but the earliest definitive remains of the other division of modern bony fish radiation — actinopterygians — are from the late Early or the earliest Middle Devonian, some 30 million years later[37]. Some scales and other skeletal detritus of late Silurian–Early Devonian age (about 427–400 Ma) are conventionally aligned with actinopterygians[38,39]. However, many — or perhaps all — of these taxa could represent stem osteichthyans[40,41] or even stem gnathostomes[42] (Fig. 3). As with chondrichthyans, early osteichthyans show some striking distributional patterns, including the conspicuous concentration of early members of major lobe-fin lineages in the latest Silurian and earliest Devonian of the South China Block[30] (Fig. 1). Outside of this restricted area, coeval bony fishes are limited to a handful of mostly fragmentary examples.

Several extinct groups join the familiar modern jawed vertebrate lineages. Armoured jawless fishes (ostracoderms) that are most often implicated as a jawed vertebrate sister group include: Middle Ordovician–Late Devonian (467–370 Ma) thelodonts, encompassing dorsoventrally flattened to cigar-shaped to deep-bodied forms[43] and bearing a shark-like shagreen of tiny scales; galeaspids, which are bottom-dwelling early Silurian–Late Devonian (439–370 Ma) fishes with flattened headshields that assume a bewildering variety of shapes and are found only in Chinese and Vietnamese deposits[44,45]; and osteostracans, which are another benthic group with spade-shaped headshields and are restricted to the middle Silurian–Late Devonian (433–372 Ma) of today's northern landmasses[45,46]. Two extinct jawed groups join this ostracoderm parade: placoderms, which are a species-rich and anatomically heterogeneous early Silurian–Late Devonian (435–360 Ma) assemblage characterized by

heavy head and trunk armour and bony jaw plates[47]; and acanthodians, which are covered in tiny scales and bear well-developed spines along the leading edges of nearly all of their fins[10] that together inspire the moniker 'spiny sharks'. The earliest fossils associated with acanthodians are isolated scales from the latest Ordovician (around 444 Ma)[27]. More reliable remains are Silurian in age, with the group's record extending to early Permian deposits (about 295 Ma) that yield the best-known and last-surviving genus *Acanthodes*[48,49].

The evolution of gnathostome phylogeny

The current picture of Palaeozoic gnathostome relationships is the product of three phases of study. Throughout, researchers have benefitted from high-quality data, thanks to the early application of physical tomography by Stensiö and the 'Stockholm school'[50–52], followed by the maturation of acid-preparation techniques in the middle of the twentieth century[53–56] and the non-destructive computed tomography of the past 15 years[42,57–60].

The modern phase of research into gnathostome relationships began with the introduction of phylogenetic systematics to vertebrate palaeontology, which had previously focused on linking species from successive geological strata as an approximate ancestor–descendant chain. Monophyly of the major taxonomic divisions of early gnathostomes was assumed, and their relative relationships were largely inferred using evidence from European and North American fossils. Within a decade of the initial application of cladistics to early vertebrates, an imperfect consensus emerged that acanthodians were a clade of stem osteichthyans[48] and that placoderms were the immediate sister group of crown gnathostomes[61]. This framework would persist for more than 30 years[10], despite the intervening discovery and detailed description of fossils from Australia[53,56,62], China[30,63] and northern Canada[64] that provided fresh morphological information beyond the stagnating stable of classic Euramerican taxa.

The second phase began in the 1980s with a cladistic reinterpretation of the ostracoderms. Detailed anatomical reinvestigations of ostracoderm sublineages and numerical phylogenetic analysis resulted in the recognition of this assemblage as a paraphyletic gnathostome stem group[65–69]. Reconfiguration of the agnathan menagerie permitted reconstructions of evolutionary patterns in fin morphology and skeletal hard tissues, and identified the extinct jawless sister group of jawed vertebrates. Although many ostracoderm lineages have been considered contenders for this position, anatomical evidence overwhelmingly supports osteostracans. Like jawed vertebrates, but unlike other agnathans, osteostracans bear well-developed pectoral fins with associated girdles, a epicercal tail, and perichondral and cellular bone (Box 1).

The third and ongoing phase is the detailed scrutiny of the pioneering cladistic framework relating acanthodians and placoderms to modern jawed vertebrate lineages. Traction on this problem arose indirectly, beginning around the turn of the century with the development of expanded numerical phylogenetic analyses targeting relationships within osteichthyans[70–73] and chondrichthyans[74–76], but employing acanthodian and placoderm outgroups. These studies introduced the use of increasingly larger data sets, and provided the character information that would seed analyses targeting not individual lineages, but early jawed vertebrates

as a whole. At the same time, a series of new fossil discoveries (outlined later) revealed unexpected anatomical combinations that raised serious questions about the coherence of acanthodians and placoderms. This set in motion a series of refined analyses of early jawed vertebrates bent on testing the supposed monophyly of these groups[42,49,58,77,78]. This final phase is a current debate and the setting for the following discussion.

New fossil discoveries and their importance

In this section, we highlight key finds since the 1980s that have challenged embedded perceptions and explain their importance in light of what is or was known about early jawed vertebrate evolution. Presented in approximate phylogenetic order, ascending from jawless members of the stem lineage, to placoderms, to members of the gnathostome crown, these discoveries provide a broad summary of the emerging picture of major evolutionary patterns in early gnathostomes. Detailed accounts of character transformation are provided elsewhere[20].

Shuyu and *Romundina* and their noses for success

The neurocranium, or braincase, is a primitively cartilaginous structure that houses the brain and paired sensory organs in vertebrates. When coated with a mineralized rind, structurally complex braincases can be
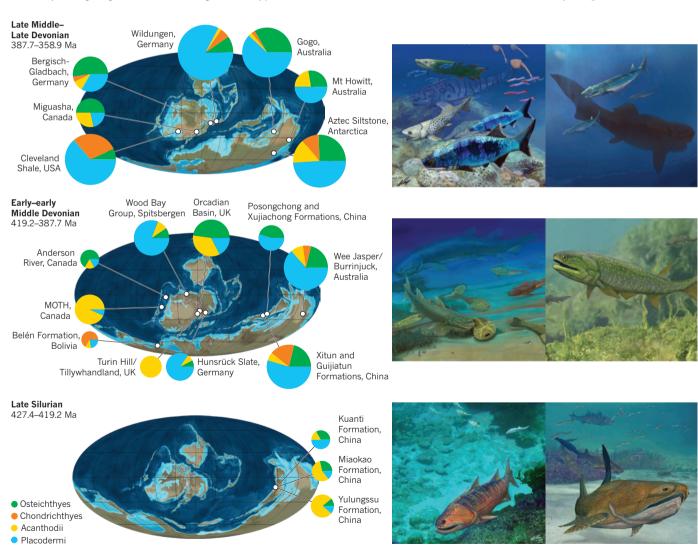
Figure 1 | Fossils relevant to early jawed-vertebrate evolution derive from major fossil sites in North America and Europe, and increasingly China and Australia. Palaeogeographic positions of localities bearing early jawed vertebrates and characterized by abundant fossils, high-fidelity preservation or both. Taxonomic breakdown of gnathostome diversity within sites is indicated by the associated pie charts and size-scaled to reported species richness.

MOTH, Man on the Hill. The vignettes depict scenes based on key fossil sites: Gogo, Australia (left) and Cleveland Shale, USA (right) in the late Middle–Late Devonian; the Xitun Formation, China (left) and Orcadian Basin, UK (right) in the Early–early Middle Devonian; and the late Silurian Kuanti Formation, China (left and right). Illustrations by B. Choo, Flinders University. Palaeogeographic reconstructions by R. Blakey, Colorado Plateau Geosystems.
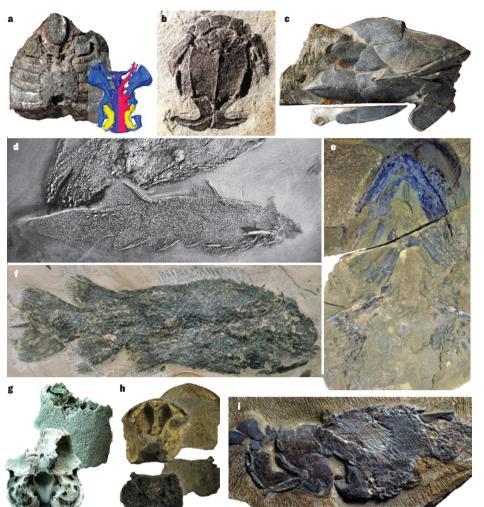
Figure 2 | Discoveries over the past two decades provide new clues about the evolution of early jawed vertebrates and their kin. a, High-fidelity virtual models of the Silurian galeaspid *Shuyu* reveal cranial architecture in jawless relatives of jawed vertebrates. **b**, Claspers in most placoderm groups, including antiarchs like *Microbrachius* shown here, raise questions about placoderm relationships and the evolution of vertebrate reproductive strategies. **c**, Osteichthyan-like pattern of bones in the Silurian placoderm *Entelognathus* suggest that the last common ancestor of all modern jawed vertebrates was clad in a bony-fish-like skeleton. **d**, Stunningly preserved fossils from the Early Devonian Man on the Hill (MOTH) locality of Canada challenges acanthodian monophyly, suggesting affinities with chondrichthyans. **e**, Pectoral-fin spines and tooth whorls with fused bases in the Early Devonian chondrichthyan *Doliodus* are features typically associated with acanthodians. **f**, The Early Devonian osteichthyan *Dialipina* shows a puzzling combination of traits despite being initially identified as a ray-finned fish based on isolated scales. **g**, An Early Devonian braincase attributed to the osteichthyan *Ligulalepis* shows features generally associated with placoderms and chondrichthyans. **h**, Braincase of *Psarolepis*, an Early Devonian lobe-finned osteichthyan from China represented by isolated bones, including spines of the kind associated with chondrichthyans, placoderms and acanthodians. **i**, The surprising reconstruction of *Psarolepis* was corroborated by the discovery of the more complete and even more ancient *Guiyu*, from the late Silurian of China. Images courtesy of **a**, Z. Gai; **b, g**, J. Long; **c, i**, M. Zhu; **e**, R. Miller; **f**, S. Cumbaa.

preserved as fossils and are a key source of phylogenetic information. Discriminating between specialized and primitive features in jawed vertebrates demands comparison with jawless fishes, but knowledge of the internal anatomy in ostracoderm lineages that lack endoskeletal mineralization is rudimentary[10,65]. By contrast, a thin coat of bone surrounds the cartilage forming the consolidated braincase and supports for the gills and pectoral fins of osteostracans. This permitted the first detailed reconstructions of osteostracan brains, cranial vessels and nerves nearly a century ago[50,51]. Galeaspids also bear a mineralized endoskeleton, but interpretations of their neurocranial structure have long been sketchy. High-resolution synchrotron scanning of the early galeaspid *Shuyu*[57] reinforced past identifications of widely separated, anterolaterally placed nasal capsules[68,79] that open medially into a central, dorsally directed duct that is also joined by the hypophysis (Fig. 2a). Thus, galeaspids show a tantalizing mosaic of cyclostome-like (nasal capsules located well behind the front of the head and opening into a common nasohypophyseal duct) and crown gnathostome-like (broad separation of nasal capsules) traits in the anterior region of the skull, and suggest that the cyclostome-like geometry of the better known osteostracans might be secondary. These features are more than just anatomical arcana — broad separation of nasal capsules is interpreted as a developmental necessity for the origin of jaws because the median nasohypophyseal placode of cyclostomes obstructs anterior growth of neural crest cells that contribute substantially to mandibles[24,57,80]. It seems that restructuring of the anterior portion of the head continued after the origin of jaws. Posteriorly placed, separate nasal capsules resembling those of galeaspids characterize the least crownward placoderms such as antiarchs, *Brindabellaspis* and *Romundina*, but these share with other jawed vertebrates a hypophysis that opens into the mouth, rather than a common nasohypophyseal duct as in agnathans[58].

By contrast, more crownward placoderms such as arthrodires, with their anteriorly placed nasal capsules, broadly resemble crown gnathostomes. These major architectural changes reflect a key piece of evidence for placoderm paraphyly[49,58,73,77,78], but ambiguities in the relationships among placoderms do not provide a consistent picture for the evolution of skull geometry in this crownward segment of the gnathostome stem.

Claspers and their evolutionary implications

The ptyctodontid placoderms have long been known to possess claspers[81], intromittent organs associated with the pelvic fins and evidence of internal fertilization. This trait factored in early cladistic investigations of placoderm intra- and interrelationships, tying placoderms to chondrichthyans[52] and fuelling arguments that ptyctodonts are the sister group of all other placoderms[10]. The discovery of arthrodire embryos within adult specimens prompted renewed investigation of this group in which long-overlooked evidence of claspers was finally discovered[82–84], followed by the realization that antiarchs also possessed these structures[85] (Fig. 2b). The palaeobiological and reproductive importance of claspers has been well considered[83,85], but their full phylogenetic importance is unresolved. Current phylogenetic consensus does not regard placoderm and chondrichthyan claspers to be homologous[20], but the homology of claspers within placoderms seems likely. Placoderm paraphyly demands the loss of internal fertilization before the origin of crown gnathostomes, signalling an unprecedented shift in reproductive biology within vertebrates[85]. Thus, we face two problematic alternatives: either internal fertilization was lost in a crownward segment of the gnathostome stem, defying observational data on the reproductive biology of living vertebrates[85], or placoderms with claspers form a clade, contradicting the apparent signal of other traits[58].

Entelognathus reframes ancestral conditions

The perceived 'primitiveness' of chondrichthyan anatomy entrenched in many general introductions to vertebrate biology has deep pre-Darwinian roots. Faced only with living species, this view seems reasonable enough: with their shagreen of tiny scales and cartilaginous internal skeletons, chondrichthyans seem to be tailor-made morphological intermediates between the naked hagfishes and lampreys on the one hand and the internally and externally bony osteichthyans on the other. The fossil record subverts this tidy picture by showing that both large dermal plates and a bony internal skeleton are innovations that arose long before the divergence of osteichthyans and chondrichthyans[22,66,67,69,86]. However, the condition of the skeleton in the last common ancestor of jawed vertebrates has remained controversial thanks to two mutually reinforcing phenomena: a reluctance to make explicit comparisons between the bony plates of osteichthyans and placoderms, and repeated interpretations of at least some acanthodians as early osteichthyan relatives[41,48,49,73,77]. Together these factors paint a picture of an ancestral crown gnathostome covered in a 'micromeric' outer skeleton of tiny scales, with a 'macromeric' skeleton composed of large plates reappearing in the osteichthyan lineage. This view was turned on its head by the discovery of the late Silurian *Entelognathus* in China[78] (about 423 Ma; Fig. 2c). Although *Entelognathus* broadly resembles a standard-issue placoderm, its cheek and upper and lower jaws are covered with bones that match the pattern seen in osteichthyans, rather than other placoderms. This remarkable correspondence suggests that there is evolutionary continuity between the large dermal plates of placoderms and those of bony fishes[42,58,78].

Man on the Hill brings acanthodians into the light

The Man on the Hill (MOTH) locality in the Northwest Territories of Canada is an Early Devonian (about 419 Ma) Konservat Lagerstätte yielding articulated early vertebrates. Originally discovered in the 1970s[64], new collections and advances in chemical preparation have since revealed exquisitely preserved fossils (Fig. 2d). Jawed vertebrates from MOTH are mostly acanthodians (Fig. 1), providing important anatomical detail on this enigmatic assemblage. Previously, the record of complete acanthodian fossils was mostly restricted to crudely prepared specimens from low-diversity, fluvial–lacustrine Early Devonian deposits of the United Kingdom[12]. By contrast, acid-prepared acanthodians from the species-rich marine MOTH locality reveal crisp anatomical details. In particular, a host of these species have umbellate and denticle-like scales such as those found in chondrichthyans[87–90]. Perhaps more importantly, the MOTH fauna include examples of acanthodian-like fishes covered in scales with growth patterns and structure previously known only from isolated fragments, but conventionally assigned to chondrichthyans[89]. This simultaneously suggests a position for acanthodians in the jawed vertebrate tree, while undermining confidence that they comprise a natural group.

The inside story on acanthodian morphology

Several early placoderms, osteichthyans and chondrichthyans yield detailed braincases[10,52,91], but acanthodian examples are rare. Subject to many re-interpretations over the past 100 years[12,48,49], the neurocranium of the Permian *Acanthodes* is central to debates on the evolutionary affinities of acanthodians. Various authors have been impressed by what they perceived as either particularly osteichthyan-like[41,48,77] or chondrichthyan-like[49,52] features of *Acanthodes*, triggering contrasting views on the placement of acanthodians as a whole. The Early Devonian (around 419 Ma) *Ptomacanthus* also preserves a braincase, although detail is obscure to the degree that this structure was initially ignored. Re-examination of *Ptomacanthus* revealed a neurocranium with a gross architecture that is more similar to that of placoderms or chondrichthyans than that of *Acanthodes* and osteichthyans, providing evidence in the first explicit argument for acanthodian paraphyly[77].

A sneak peek at early shark anatomy

With a sparse early record, interpretation of primitive chondrichthyan conditions drew heavily on body fossils from the latest Devonian[11] and even younger braincases[91], all of which are probably highly specialized. This changed with two stunning finds in the early 2000s. First was the discovery of more complete neurocrania of *Pucapampella* from the Early Devonian of Bolivia[76] and a similar South African form[92]. Previously named on the basis of an isolated neurocranial base, *Pucapampella* bears a chondrichthyan-specific hard tissue (prismatic calcified cartilage) in combination with a ventral fissure: a persistent division between two embryonic braincase components. Absent in ostracoderms, placoderms and other chondrichthyans, but present in *Acanthodes* and bony fishes, the ventral fissure was long considered key evidence for a close relationship between acanthodians and osteichthyans[48]. *Pucapampella* suggests that this trait is a general feature of crown-group gnathostomes. Subsequent discoveries provided additional anatomical details for *Pucapampella*, revealing peculiar teeth and jaws to accompany its unanticipated neurocranial architecture[34]. Hot on the heels of *Pucapampella* came the discovery of the oldest articulated chondrichthyan. *Doliodus*, from the Early Devonian of New Brunswick[93], was known for more than a century only by isolated teeth, and assigned to acanthodians. Recovery of an articulated head and forequarters revealed the signature chondrichthyan trait of prismatic calcified cartilage occurring in a fish with stubby spines along the leading edges of its pectoral fins (Fig. 2e), casting further doubt on acanthodian monophyly. Subsequent analysis of the braincase[59] and dentition[60,94] of *Doliodus* revealed primitive character states, such as fused tooth bases, not widely seen in crown chondrichthyans and certainly absent in modern sharks and rays, but common to acanthodians and early osteichthyans.

Rosetta stones for fragmentary bony fish remains

Fossil bony fishes have conventionally been deposited in one of the two living divisions: actinopterygians or sarcopterygians. This leaves the osteichthyan stem bereft of fossils that document the origin of this enormously successful clade. A series of isolated scales of late Silurian–Early Devonian age were loosely tethered to actinopterygians as their representatives[38,39], but the discovery of more complete material attributed to *Dialipina*[95] (Fig. 2f) and *Ligulalepis*[54,55] (Fig. 2g) raised questions about their actinopterygian affinities, and the importance of scale-based characters used to identify ray-finned fishes[41,73]. The braincase aligned with the scale-taxon *Ligulalepis* shows evidence of an eyestalk[54,55], a cartilaginous plinth that supports the eye in chondrichthyans and placoderms, but that is absent in modern osteichthyans. This might suggest *Ligulalepis* is a stem osteichthyan, but reports of eyestalks in early sarcopterygians[72] argue for parallel loss in the two bony fish divisions. Complete specimens of *Dialipina* are even more puzzling, marrying a tail geometry found only in lobe-finned fishes with a cheek comprising tiny bones that bear no clear resemblance to the large plates of other osteichthyans or even *Entelognathus*. *Ligulalepis* and *Dialipina* vacillate between Actinopterygii and the osteichthyan stem in many analyses[42,78], and solid placements are likely to be elusive until these taxa are more completely documented.

Psarolepis and *Guiyu* encapsulate the revolution

Perhaps more than any other discovery, *Psarolepis* represents the principal instigator of the current revolution in early jawed-vertebrate systematics. Recovered from late Silurian and earliest Devonian rocks of China, it is one of the earliest bony fishes (Fig. 1). First identified as a stem lungfish on the basis of jaw and braincase material[96], subsequent investigation of *Psarolepis* and the discovery of isolated cheek and shoulder bones highlighted more interesting affinities[70]. *Psarolepis* exhibits two hallmarks of the lobe-finned fishes: a braincase divided into front and hind units by an articulating joint and a pore-canal complex in its dermal bones (Fig. 2h). However, the cleaver-shaped cheek and maxilla (upper external jaw bone) bear an uncanny resemblance to those of early ray-finned fishes, suggestive of a shared primitive condition for bony fishes. More surprisingly, *Psarolepis* bristled with spines: the shoulder girdle bears a pronounced spine over the fin articulation area, reminiscent of acanthodians and some placoderms, whereas the dorsal fins were preceded by spines like those of chondrichthyans and acanthodians. *Psarolepis* is most reasonably interpreted as a
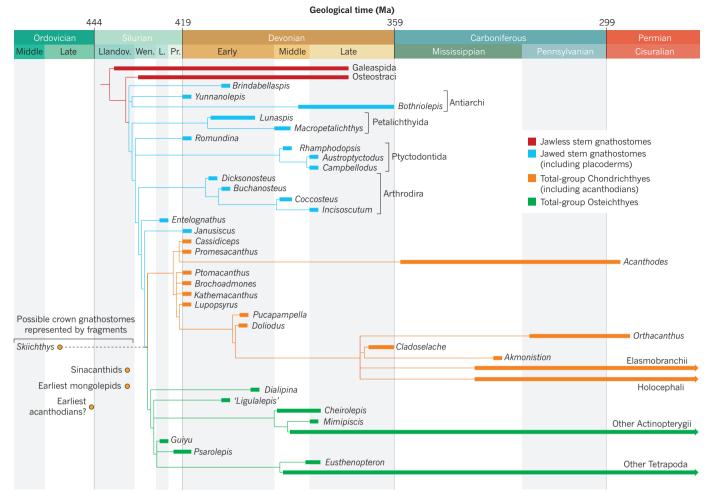
Figure 3 | Time-calibrated phylogeny of early jawed vertebrates and their immediate jawless relatives, showing minimum times of divergence based on fossil evidence. Topology based on ref. 42, with some taxa omitted for clarity and modifications showing presumed phylogenetic positions of key extant lineages. Also shown are key early jawed vertebrates or putative jawed vertebrates with uncertain affinities to the crown group. The minimum age of the gnathostome crown could be profoundly recalibrated if *Skiichthys*[28] is confirmed as a crown-group gnathostome. Dotted line indicates possible range extension for the gnathostome crown node. Llandov., Llandovery; Wen. Wenlock; L., Ludlow; Pr., Přídolí.

stem-group sarcopterygian[9,72,73], and thus an early example of the bony fish lineage that would give rise to tetrapods. However, it is held in this position by such a small number of traits, and retains so many plesiomorphies, that some analyses have recovered it as a stem-group osteichthyan[70,71]. This shook confidence in the seemingly stable, decades-old sets of attributes that characterize major early vertebrate groups[10]. However, the disarticulated nature of these fossils raised the troubling possibility that the combination of characters in *Psarolepis* was chimaeric: parts of different species misattributed to a single one. This concern was rejected, albeit indirectly, by the discovery of *Guiyu*[9] (Fig. 2j). Broadly similar to *Psarolepis,* but from even older Silurian rocks in China (about 423 Ma), *Guiyu* provides exceptional corroboration that traits such as a jointed braincase occurred in the same animal as pectoral- and dorsal-fin spines, and delivers further surprises, including the presence of placoderm-like external pelvic girdles[97]. Interpreted as an early sarcopterygian, *Guiyu* also shows that the last common ancestor of all modern osteichthyans arose no later than the Silurian, before the Devonian 'Age of Fishes'.

The re-shaping of early jawed vertebrate phylogeny
This panoply of new taxa and unexpected character distributions fuelled doubts about the status of classic early jawed vertebrate catagories[93,98], but early studies did not match these queries with cladistic tests. In the past five years, the field has witnessed a spate of numerical analyses giving rise to rapidly shifting perspectives on phylogenetic relationships[9,42,49,58,77,85]. However, some stable patterns are apparent and key areas of ongoing

debate are now coming into focus.

The monophyly of fossil osteichthyans and chondrichthyans is universally supported. Placoderms are repeatedly recovered as stem-group gnathostomes and acanthodians are generally agreed to be members of the gnathostome crown, with some noteworthy exceptions[49]. Major differences with previous hypotheses stem from important shifts in approach, such as abandoning earlier assumptions of placoderm and acanthodian monophyly. In all cases so far, the monophyly of placoderms has been rejected and, in all but one[58], acanthodian monophyly has also been rejected.

In the earliest iterations, acanthodians were inferred to be massively paraphyletic, with some members associated with chondrichthyan, osteichthyan and gnathostome stem branches[49,58]. This configuration helped to explain the odd conjunction of osteichthyan, chondrichthyan and more primitive characters found in acanthodians. Furthermore, it implied an acanthodian-like appearance of the ancestral crown gnathostome: a small fusiform fish, covered in a denticle shagreen, a skull composed of mostly undifferentiated plates, with spines preceding the fins. The unfortunate complication of this hypothesis was that it implied nonhomology of osteichthyan and placoderm armoured exoskeletons. Similarities between osteichthyan and placoderm skulls and shoulder girdles had not gone unnoticed[70,71,99], but were matched by dismissals citing 'fundamental differences' in construction[100]. The discovery of *Entelognathus* (already discussed) deals a blow to the latter perspective. Phylogenetic analysis accompanying the discovery[78] unsurprisingly led to a wholesale

shift of acanthodian-type taxa to the chondrichthyan total group. Every subsequent analysis has corroborated this outcome[42,58,85]. This key rearrangement eliminates the need to invoke convergence between placoderm and osteichthyan exoskeletons. By viewing the fragmented dermal skeletons of chondrichthyans and acanthodians as a derived condition, no special sister group relationship between osteichthyans and placoderms is implied, as had been assumed in the past[99].

Current analyses universally reject placoderm monophyly, with arthrodires (and similar forms such as *Entelognathus*) resolved closest to the gnathostome crown (Fig. 3). This arrangement suggests that resemblances between arthrodires and modern gnathostomes are homologous — a point reinforced by the arthrodire gestalt of *Entelognathus*. Likewise, it suggests that the similarities between the more flat-headed and presumably benthic placoderms, such as antiarchs and petalichthyids, and jawless outgroups reflect a shared primitive condition[10,20,77]. This has the convenient effect of stretching the placoderms into an array of jaw-bearing stem gnathostomes, although mandibles remain unknown in forms such as *Brindabellaspis* and petalichthyids.

The consistency of placoderm paraphyly across recent analyses[20,42,49,58,73,77,78,85] suggests that this is well supported. However, available solutions are not wholly independent, with each data set incrementally updated from a core original study[77]. Perhaps notably, the addition of taxa and characters has not increased support for the paraphyletic placoderm backbone. Instead, successive analyses have seen a winnowing of branch support for the deepest divergences among jaw-bearing stem gnathostomes, coupled with inconsistent arrangements of major placoderm lineages crownward of antiarchs and *Brindabellaspis*. This instability, combined with potential placoderm synapomorphies such as pelvic claspers[85] and a persistent fissure between the nasal capsules and the remainder of the braincase[20], indicate that the 'placoderm problem' is far from resolved. A satisfactory resolution of the relationships of placoderms will have profound consequences for our understanding of the origin of modern jawed vertebrates.

Future directions

Early jawed vertebrate phylogenetics is in a state of infancy, but rapid progress is being made. Present discourse on early jawed vertebrate phylogenetics is marked by a growth of healthy debate and a relative lack of the kind of dogmatism that held back the field for nearly half a century. The question of the origin of the jaws themselves remains open. So far, the problem has been debated in terms of highly idealized archetypal scenarios, such as the transformation of gill arches into jaws[1]. From both palaeontological and neontological perspectives, this scenario has proved deficient[6,10,80]. Little direct evidence of the visceral skeleton of fossil jawless fishes is known; even the proximate outgroups of the jawed vertebrates — osteostracans and galeaspids — are presumed to have been jawless, but remains of the oral skeleton remain absent. What is known of the oral regions of osteostracans and galeaspids suggests that they possessed mouths that were specialized relative to the branchial arches, a condition consistent with modern jawless fishes[10]. Placoderm paraphyly raises some hope that relevant data could be sourced from this assemblage (for example, *Brindabellaspis* or petalichthyids). The discovery of additional fossils will hopefully help to fill these gaps, but they will not be sufficient by themselves. Rigorous phylogenetic analysis must accompany these new finds to avoid simply shoehorning fossils into appealing narratives[12]. ∎

1. Gegenbaur, C., Bell, F. J. & Lankester, E. R. *Elements of Comparative Anatomy* (Macmillan and Co., 1878).
2. Balfour, F. M. On the development of the skeleton of the paired fins of Elasmobranchii, considered in relation to its bearings on the nature of the limbs of the Vertebrata. *Proc. Zool. Soc. Lond.* **49,** 656–670 (1881).
3. de Beer, G. *The Development of the Vertebrate Skull* (Oxford Univ. Press, 1937).
4. Reif, W.-E. Evolution of dermal skeleton and dentition in vertebrates. *Evol. Biol.* **15,** 287–368 (1982).
5. Shubin, N. H. Origin of evolutionary novelty: examples from limbs. *J. Morphol.* **252,** 15–28 (2002).
6. Kuratani, S. Evolution of the vertebrate jaw: comparative embryology and
molecular developmental biology reveal the factors behind evolutionary novelty. *J. Anat.* **205,** 335–347 (2004).
7. Shigetani, Y., Sugahara, F. & Kuratani, S. A new evolutionary scenario for the vertebrate jaw. *Bioessays* **27,** 331–338 (2005).
8. Wagner, G. P. & Lynch, V. J. Evolutionary novelties. *Curr. Biol.* **20,** R48–R52 (2010).
9. Zhu, M. et al. The oldest articulated osteichthyan reveals mosaic gnathostome characters. *Nature* **458,** 469–474 (2009).
10. Janvier, P. *Early Vertebrates* (Clarendon, 1996). **This masterful summary provides a window on the 'state of the art' immediately preceding the major changes to our understanding of relationships among early gnathostomes that took place over the past two decades, and is still an indispensible and accessible resource.**
11. Dean, B. Contributions to the morphology of *Cladoselache* (*Cladodus*). *J. Morphol.* **9,** 87–114 (1894).
12. Watson, D. M. S. The acanthodian fishes. *Philos. Trans. R. Soc. Lond.* **228,** 49–146 (1937).
13. Zangerl, R. & Williams, M. E. New evidence on the nature of the jaw suspension in Palaeozoic anacanthous sharks. *Palaeontology* **18,** 333–341 (1975).
14. Gregory, W. K. Further observations on the pectoral girdle and fin of *Sauripterus taylori* Hall, a crossopterygian fish from the Upper Devonian of Pennsylvania, with special reference to the origin of the pentadactylate extremities of Tetrapoda. *Proc. Am. Phil. Soc.* **75,** 673–690 (1935).
15. Miles, R. S. A reinterpretation of the visceral skeleton of *Acanthodes*. *Nature* **204,** 457–459 (1964).
16. Davis, M. C., Shubin, N. & Daeschler, E. B. A new specimen of *Sauripterus taylori* (Sarcopterygii, Osteichthyes) from the Famennian Catskill Formation of North America. *J. Vertebr. Paleontol.* **24,** 26–40 (2004).
17. Kemp, T. S. *The Origin and Evolution of Mammals* (Oxford Univ. Press, 2005).
18. Makovicky, P. J. & Zanno, L. E. in *Living Dinosaurs: The Evolutionary History of Modern Birds* (eds Dyke, G. & Kaiser, G.) 9–29 (Wiley, 2011).
19. Clack, J. A. *Gaining Ground* (Indiana Univ. Press, 2012).
20. Brazeau, M. D. & Friedman, M. The characters of Palaeozoic jawed vertebrates. *Zool. J. Linn. Soc.* **170,** 779–821 (2014).
21. Chen, M., Zou, M., Yang, L. & He, S. Basal jawed vertebrate phylogenomics using transcriptomic data from Solexa sequencing. *PLoS ONE* **7,** e36256 (2012).
22. Donoghue, P. C., Forey, P. L. & Aldridge, R. J. Conodont affinity and chordate phylogeny. *Biol. Rev. Camb. Philos. Soc.* **75,** 191–251 (2000).
23. Ota, K. G., Fujimoto, S., Oisi, Y. & Kuratani, S. Identification of vertebra-like elements and their possible differentiation from sclerotomes in the hagfish. *Nature Commun.* **2,** 373 (2011).
24. Oisi, Y., Ota, K. G., Kuraku, S., Fujimoto, S. & Kuratani, S. Craniofacial development of hagfishes and the evolution of vertebrates. *Nature* **493,** 175–180 (2013).
25. Ota, K. G., Kuraku, S. & Kuratani, S. Hagfish embryology with reference to the evolution of the neural crest. *Nature* **446,** 672–675 (2007).
26. Heimberg, A. M., Cowper-Sal-lari, R., Semon, M., Donoghue, P. C. & Peterson, K. J. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc. Natl Acad. Sci. USA* **107,** 19379–19383 (2010).
27. Karatajute-Talimaa, V. & Predtechenskyj, N. The distribution of the vertebrates in the Late Ordovician and Early Silurian palaeobasins of the Siberian Platform. *Bull. Mus. Natl Hist. Nat.* **4,** 39–55 (1995).
28. Smith, M. M. & Sansom, I. J. Exoskeletal micro-remains of an Ordovician fish from the Harding Sandstone of Colorado. *Palaeontology* **40,** 645–658 (1997).
29. Sansom, I. J., Davies, N. S., Coates, M. I., Nicoll, R. S. & Ritchie, A. Chondrichthyan-like scales from the Middle Ordovician of Australia. *Palaeontology* **55,** 243–247 (2012).
30. Zhao, W.-J. & Zhu, M. Siluro-Devonian vertebrate biostratigraphy and biogeography of China. *Palaeoworld* **19,** 4–26 (2010).
31. Anderson, P. S., Friedman, M., Brazeau, M. D. & Rayfield, E. J. Initial radiation of jaws demonstrated stability despite faunal and environmental change. *Nature* **476,** 206–209 (2011).
32. Karatajute-Talimaa, V. N., Novtistkaya, L. I., Rozman, K. S. & Sodov, J. *Mongolepis,* a new genus of Elasmobranchii from the Lower Silurian of Mongolia. *Paleontologicheskii zhurnal* **1,** 76–86 (1990).
33. Sansom, I. J., Wang, N.-Z. & Smith, M. The histology and affinities of sinacanthid fishes: primitive gnathostomes from the Silurian of China. *Zool. J. Linn. Soc.* **144,** 379–386 (2005).
34. Janvier, P. & Maisey, J. G. in *Morphology, Phylogeny and Paleobiogeography of Fossil Fishes* (eds Elliott, D. K., Maisey, J. G., Yu, X. & Miao, D.) 431–459 (Dr Freidrich Pfeil, 2010).
35. Panchen, A. L. & Smithson, T. R. Character diagnosis, fossils and the origin of tetrapods. *Biol. Rev. Camb. Philos. Soc.* **62,** 341–436 (1987).
36. Ahlberg, P. E. & Johanson, Z. Osteolepiforms and the ancestry of tetrapods. *Nature* **395,** 792–794 (1998).
37. Lukševičs, E., Lebedev, O. A. & Zakharenko, G. V. Palaeozoogeographical connections of the Devonian vertebrate communities of the Baltica Province. Part I. Eifelian-Givetian. *Palaeoworld* **19,** 94–107 (2010).
38. Schultze, H.-P. Palaeoniscoidea-Schuppen aus dem Unterdevon Australiens und Kanadas und aus dem Mitteldevon Spitzbergens [in German]. *British Mus. Nat. Hist. Geol.* **16,** 343–376 (1968).
39. Gross, W. Fragliche Actinopterygier-Schuppen aus dem Silur Gotlands [in German]. *Lethaia* **1,** 184–218 (1968).
40. Botella, H., Blom, H., Dorka, M., Ahlberg, P. E. & Janvier, P. Jaws and teeth of the earliest bony fishes. *Nature* **448,** 583–586 (2007).
41. Friedman, M. & Brazeau, M. D. A reappraisal of the origin and basal radiation of the Osteichthyes. *J. Vertebr. Paleontol.* **30,** 36–56 (2010).

42. Giles, S., Friedman, M. & Brazeau, M. D. Osteichthyan-like cranial conditions in an Early Devonian stem gnathostome. *Nature* http://dx.doi.org/10.1038/nature14065 (2015).
43. Märss, T., Turner, S. & Karatajute-Talimaa, V. in *Handbook of Paleoichthyology* Vol. 1B (ed. Schultze, H.-P.) (Dr Friedrich Pfeil, 2007).
44. Zhu, M. & Gai, Z.-K. Phylogenetic relationships of galeaspids (Agnatha). *Vertebr. PalAsiat.* **44,** 1–27 (2006).
45. Sansom, R. S. Endemicity and palaeobiogeography of the Osteostraci and Galeaspida: a test of scenarios of gnathostome evolution. *Palaeontology* **52,** 1257–1273 (2009).
46. Sansom, R. S. Phylogeny, classification and character polarity of the Osteostraci (Vertebrata). *J. Syst. Paleontol.* **7,** 95–115 (2009).
47. Young, G. C. Placoderms (armoured fish): dominant vertebrates of the Devonian period. *Annu. Rev. Earth Planet. Sci.* **38,** 523–550 (2010).
48. Miles, R. S. in *Interrelationships of Fishes* (eds Greenwood, P. H., Miles, R. S. & Patterson, C.) 63–103 (Academic, 1973).
    This first-generation application of cladistic methodology to early jawed vertebrates placed the 'spiny sharks' as early relatives of bony fishes, a perspective that profoundly influenced perceptions of the ancestral crown gnathostome for more than 40 years.
49. Davis, S. P., Finarelli, J. A. & Coates, M. I. *Acanthodes* and shark-like conditions in the last common ancestor of modern gnathostomes. *Nature* **486,** 247–250 (2012).
50. Stensiö, E. A. The Devonian and Downtonian vertebrates of Spitsbergen. Part 1. Family Cephalaspidae. *Skr. Svalbard Ishav.* **12,** 1–391 (1927).
51. Stensiö, E. A. *The Cephalaspids of Great Britain* (British Museum (Natural History), 1932).
52. Jarvik, E. *Basic Structure and Evolution of Vertebrates* (Academic, 1980).
53. White, E. I. The larger arthrodiran fishes from the area of the Burrinjuck Dam, N.S.W. *Tran. Zoo. Soc. Lond.* **34,** 149–262 (1978).
54. Basden, A. M. & Young, G. C. A primitive actinopterygian neurocranium from the Early Devonian of Southeastern Australia. *J. Vertebr. Paleontol.* **21,** 754–766 (2001).
55. Basden, A. M., Young, G. C., Coates, M. I. & Richtie, A. The most primitive osteichthyan braincase? *Nature* **403,** 185–188 (2000).
56. Young, G. C. A new Early Devonian placoderm from New South Wales, Australia, with a discussion of placoderm phylogeny. *Palaeontogr. A* **167,** 10–76 (1980).
57. Gai, Z., Donoghue, P. C., Zhu, M., Janvier, P. & Stampanoni, M. Fossil jawless fish from China foreshadows early jawed vertebrate anatomy. *Nature* **476,** 324–327 (2011).
58. Dupret, V., Sanchez, S., Goujet, D., Tafforeau, P. & Ahlberg, P. E. A primitive placoderm sheds light on the origin of the jawed vertebrate face. *Nature* **507,** 500–503 (2014).
59. Maisey, J. G., Miller, R. & Turner, S. The braincase of the chondrichthyan *Doliodus* from the Lower Devonian Campbellton Formation of New Brunswick, Canada. *Acta Zool.* **90** (Suppl. 1), 109–122 (2009).
60. Maisey, J. G., Turner, S., Naylor, G. J. & Miller, R. F. Dental patterning in the earliest sharks: implications for tooth evolution. *J. Morphol.* **275,** 586–596 (2014).
61. Schaeffer, B. in *Problèmes Actuels de Paléontologie: Evolution des Vertébrés* Vol. 218 [in French] (ed. Lehman, J.-P.) 101–109 (Colloques internationaux du Centre national de la Recheche scientifique, 1975).
62. Long, J. A. & Trinajstic, K. The Late Devonian Gogo Formation Lägerstatte of Western Australia: exceptional early vertebrate preservation and diversity. *Annu. Rev. Earth Planet. Sci.* **38,** 255–279 (2010).
63. Zhu, M. Catalogue of Devonian vertebrates in China, with notes on bio-events. *Cour. Forsch. Inst. Senckenberg* **223,** 379–390 (2000).
64. Bernacsek, G. M. & Dineley, D. L. New acanthodians from the Delorme Formation (Lower Devonian) of N.W.T. Canada. *Palaeontogr. A* **159,** 1–25 (1977).
65. Janvier, P. & Blieck, A. New data on the internal anatomy of the Heterostraci (Agnatha), with general remarks on the phylogeny of the Craniota. *Zool. Scr.* **8,** 287–296 (1979).
66. Janvier, P. The phylogeny of Craniata, with particular reference to the significance of fossil 'agnathans'. *J. Vertebr. Paleontol.* **1,** 121–159 (1981).
    This article established osteostracans and galeaspids as successive outgroups to, and thus important comparative models for, jawed vertebrates, an arrangement that has survived intact for more than three decades.
67. Forey, P. L. Yet more reflections on agnathan-gnathostome relationships. *J. Vertebr. Paleontol.* **4,** 330–343 (1984).
68. Wang, N.-Z. in *Early Vertebrates and Related Problems of Evolutionary Biology* (eds Chang, M.-M., Lui, Y.-H. & Zhang, G.-R.) (Science, 1991).
69. Forey, P. L. & Janvier, P. Agnathans and the origin of jawed vertebrates. *Nature* **361,** 129–134 (1993).
70. Zhu, M., Yu, X. & Janvier, P. A primitive fossil fish sheds light on the origin of bony fishes. *Nature* **397,** 607–610 (1999).
    The bizarre combination of traits for *Psarolepis* reported in this article highlighted weaknesses in existing phylogenies of early jawed vertebrates, and triggered a resurgence in systematic studies.
71. Zhu, M. & Schultze, H.-P. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 81–84 (Taylor & Francis, 2001).
72. Zhu, M., Yu, X. & Ahlberg, P. E. A primitive sarcopterygian fish with an eyestalk. *Nature* **410,** 81–84 (2001).
73. Friedman, M. *Styloichthys* as the oldest coelacanth: implications for early osteichthyan interrelationships. *J. Syst. Palaeontology* **5,** 289–343 (2007).
74. Coates, M. I. & Sequeira, S. E. K. A new stethacanthid chondrichthyan from the Lower Carboniferous of Bearsden, Scotland. *J. Vertebr. Paleontol.* **21,** 438–459 (2001).
75. Coates, M. I. & Sequeira, S. E. K. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 241–262 (Taylor & Francis, 2001).
76. Maisey, J. G. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 263–288 (Taylor & Francis, 2001).
77. Brazeau, M. D. The braincase and jaws of a Devonian 'acanthodian' and modern gnathostome origins. *Nature* **457,** 305–308 (2009).
    This study was the first to rigorously test — and, in doing so, to reject — placoderm and acanthodian monophyly, and provides the empirical core for most subsequent phylogenetic investigations of early gnathostomes.
78. Zhu, M. *et al.* A Silurian placoderm with osteichthyan-like marginal jaw bones. *Nature* **502,** 188–193 (2013).
    Of the many remarkable early gnathostome fossils to emerge from China, few have shifted the evolutionary paradigm as much as *Entelognathus*, a placoderm-like creature with jaw bones resembling those of bony fishes.
79. Halstead, L. B. Internal anatomy of the polybranchiaspids (Agnatha, Galeaspida). *Nature* **282,** 833–836 (1979).
80. Kuratani, S. Evolution of the vertebrate jaw from developmental perspectives. *Evol. Dev.* **14,** 76–92 (2012).
81. Miles, R. S. Observations on the ptyctodont fish, *Rhamphodopsis* Watson. *Zool. J. Linn. Soc.* **47,** 99–120 (1967).
82. Ahlberg, P., Trinajstic, K., Johanson, Z. & Long, J. Pelvic claspers confirm chondrichthyan-like internal fertilization in arthrodires. *Nature* **460,** 888–889 (2009).
    This direct evidence of claspers in arthrodires renewed the palaeobiological importance of placoderms regarding internal fertilization, but potentially weakens the case for their paraphyly.
83. Trinajstic, K., Boisvert, C., Long, J., Maksimenko, A. & Johanson, Z. Pelvic and reproductive structures in placoderms (stem gnathostomes). *Biol. Rev. Camb. Philos. Soc.* http://dx.doi.org/10.1111/brv.12118 (2014).
84. Long, J. A., Trinajstic, K. & Johanson, Z. Devonian arthrodire embryos and the origin of internal fertilization in vertebrates. *Nature* **457,** 1124–1127 (2009).
85. Long, J. A. *et al.* Copulation in antiarch placoderms and the origin of gnathostome internal fertilization. *Nature* **517,** 196–199 (2015).
86. Janvier, P. The relationships of the Osteostraci and Galeaspida. *J. Vertebr. Paleontol.* **4,** 344–358 (1984).
87. Hanke, G. F. & Wilson, M. V. H. in *Recent Advances in the Origin and Early Radiation of Vertebrates* (eds Arratia, G., Wilson, M. V. H. & Cloutier, R.) 189–216 (Dr Friedrich Pfeil, 2004).
88. Hanke, G. F. & Wilson, M. V. H. in *Morphology, Phylogeny and Paleobiogeography of Fossil Fishes* (eds Elliott, D. K., Maisey, J. G., Yu, X. & Miao, D.) 149–182 (Dr Friedrich Pfeil, 2010).
89. Hanke, G. F., Wilson, M. V. H. & Saurette, F. Partial articulated specimen of the Early Devonian putative chondrichthyan *Polymerolepis whitei* Karatajūtė-Talimaa, 1968, with an anal fin spine. *Geodiversitas* **35,** 529–543 (2013).
90. Hanke, G. F. & Wilson, M. V. H. Anatomy of the Early Devonian acanthodian *Brochoadmones milesi* based on nearly complete body fossils, with comments on the evolution and development of paired fins. *J. Vertebr. Paleontol.* **26,** 526–537 (2006).
91. Schaeffer, B. The xenacanth shark neurocranium, with comments on elasmobranch monophyly. *Bull. Am. Mus. Nat. Hist.* **169,** 1–66 (1981).
92. Maisey, J. G. & Anderson, M. E. A primitive chondrichthyan braincase from the Early Devonian of South Africa. *J. Vertebr. Paleontol.* **21,** 702–713 (2001).
93. Miller, R. F., Cloutier, R. & Turner, S. The oldest articulated chondrichthyan from the Early Devonian period. *Nature* **425,** 501–504 (2003).
    This reports the oldest record of an articulated chondrichthyan and the first example with paired fin spines, initiating the dissolution of support for acanthodian monophyly.
94. Turner, S. in *Recent Advances in the Origin and Early Radiation of Vertebrates* (eds Arratia, G., Wilson, M. V. H. & Cloutier, R.) 67–94 (Dr Friedrich Pfeil, 2004).
95. Schultze, H.-P. & Cumbaa, S. L. in *Major Events in Early Vertebrate Evolution* (ed. Ahlberg, P. E.) 315–332 (Taylor & Francis, 2001).
96. Yu, X. A new porolepiform-like fish, *Psarolepis romeri*, gen. et sp. nov. (Sarcopterygii, Osteichthyes) from the Lower Devonian of Yunnan, China. *J. Vertebr. Paleontol.* **18,** 261–274 (1998).
97. Zhu, M. *et al.* Fossil fishes from China provide first evidence of dermal pelvic girdles in osteichthyans. *PLoS ONE* **7,** e35103 (2012).
98. Coates, M. I. The evolution of paired fins. *Theory Biosci.* **122,** 266–287 (2003).
99. Gardiner, B. G. The relationships of placoderms. *J. Vertebr. Paleontol.* **4,** 375–395 (1984).
100. Young, G. C. The relationships of the placoderm fishes. *Zool. J. Linn. Soc.* **88,** 1–57 (1986).
    This article provided an explicit argument for the status of placoderms as stem gnathostomes that has not been seriously challenged in the following three decades.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/snrujd. Correspondence should be addressed to M.B (m.brazeau@imperial.ac.uk) or M.F. (mattf@earth.ox.ac.uk).

ARTICLE

Oxytocin enables maternal behaviour by balancing cortical inhibition

Bianca J. Marlin[1,2,3,4], Mariela Mitre[1,2,3,4,5,6], James A. D'amour[1,2,3,4], Moses V. Chao[1,2,4,5,6,7] & Robert C. Froemke[1,2,3,4,7]

Oxytocin is important for social interactions and maternal behaviour. However, little is known about when, where and how oxytocin modulates neural circuits to improve social cognition. Here we show how oxytocin enables pup retrieval behaviour in female mice by enhancing auditory cortical pup call responses. Retrieval behaviour required the left but not right auditory cortex, was accelerated by oxytocin in the left auditory cortex, and oxytocin receptors were preferentially expressed in the left auditory cortex. Neural responses to pup calls were lateralized, with co-tuned and temporally precise excitatory and inhibitory responses in the left cortex of maternal but not pup–naive adults. Finally, pairing calls with oxytocin enhanced responses by balancing the magnitude and timing of inhibition with excitation. Our results describe fundamental synaptic mechanisms by which oxytocin increases the salience of acoustic social stimuli. Furthermore, oxytocin–induced plasticity provides a biological basis for lateralization of auditory cortical processing.

The neuropeptide oxytocin controls social behaviours such as pair bond formation, mating and parenting[1–14]. Oxytocin is synthesized in the paraventricular nucleus (PVN) and supraoptic nucleus of the hypothalamus, and binds to a G-protein-coupled receptor with a single isoform[1,2]. Peripheral release of oxytocin is important for parturition and lactation[2,7,9], whereas central release of oxytocin appears to have cognitive effects including increased interpersonal trust and enhanced salience of socially relevant sensory input[1,3,4,6,10]. However, it remains unclear which neurons express oxytocin receptors[15,16], or how oxytocin interacts with experience to modify neural circuits and increase the salience of social information.

Here we examine how oxytocin is involved in pup retrieval, an important form of mammalian social behaviour. Mouse pups emit ultrasonic distress calls when separated from the nest, which experienced mothers (known as dams) use to locate and retrieve isolated pups[17–23]. This behaviour relies on the auditory system, as pup calls played by speakers attract maternal animals[19,21]. Physiologically, neural responses to pup calls in the mouse auditory cortex differ between dams and virgin females, with higher signal-to-noise ratios in maternal mice[23–27]. Correspondingly, most inexperienced animals do not initially retrieve pups[28]. Intriguingly, some virgin female rodents start retrieving pups after being co-housed with dam and pups or after central administration of oxytocin[7]. An ethologically important form of plasticity in the auditory cortex might therefore be enabled by oxytocin in maternal animals, allowing them to recognize the behavioural significance of infant distress calls. Here we aim to show how these neural changes occur, and what role oxytocin has in experience-dependent pup retrieval by virgins.

Oxytocin enables pup retrieval behaviour

We first determined the time course of experience-dependent pup retrieval behaviour enabled by oxytocin (Fig. 1a). Non-retrieving virgin female mice were co-housed with dams and litters, and retrieval success rates of virgins were tested over 3–7 days. Mothers were first tested to ensure that they reliably retrieved pups (Fig. 1b, d and Supplementary Video 1). Three groups of virgins were examined

for pup retrieval. The first group of wild-type virgins received systemic oxytocin injections before testing (Fig. 1b–d, red). The second wild-type group received saline vehicle injections (Fig. 1b–d, black). The third optogenetic group of oxytocin-IRES-Cre mice[29,30] (*Oxt-IRES-Cre*; which express Cre recombinanse under the control of endogenous *Oxt*) expressed the channelrhodopsin-2 variant ChETA (containing a Glu123Thr mutation) in PVN oxytocin neurons, with optical fibres implanted in PVN to enhance release of endogenous oxytocin and perhaps other co-factors during retrieval testing (Fig. 1b–d, blue; Extended Data Fig. 1).
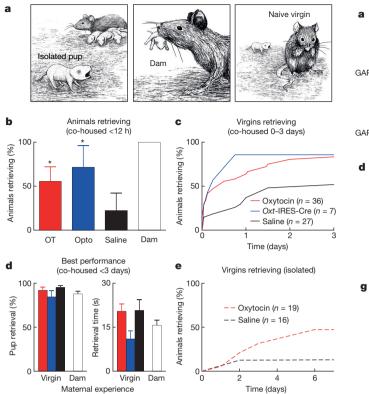
Within 12 h of being co-housed, virgin females receiving either oxytocin or optical PVN stimulation began retrieving more than saline-injected animals (Fig. 1b, c and Extended Data Fig. 2). Saline-injected virgins generally required at least 2 days of co-housing to express retrieval behaviour (Fig. 1b, c, black; Supplementary Video 2). Systemic oxytocin or optogenetic stimulation accelerated and increased retrieval, even sometimes after only a few hours of co-housing (Fig. 1b, c and Supplementary Video 3). Once retrieving, experienced virgin and dam retrieval rates and speeds were similar (Fig. 1d).

We examined retrieval in single-housed virgin females, to determine whether the effects of oxytocin required co-housing with dam and litter. Isolated virgins receiving oxytocin injections began retrieving earlier than saline-injected virgins, although slower than co-housed virgins (Fig. 1e). This demonstrates that oxytocin enables retrieval in single-housed virgins, specifically during interactions with isolated pups.

Oxytocin receptor expression is lateralized

It is unknown where in the brain oxytocin acts to improve social cognition and enable maternal behaviour. While peripheral oxytocin injections or nasal sprays have pro-social effects, it remains unclear how and where oxytocin acts on neural circuits[1,6] outside recent studies in transgenic mice[15,16]. To determine which cells express oxytocin receptors, we generated a specific oxytocin receptor antibody, OXTR-2 (Fig. 2a). The antibody labelled a subset of cells in the auditory cortex and other areas (Fig. 2b and Extended Data Fig. 3) enriched for

[1]Skirball Institute for Biomolecular Medicine, New York University School of Medicine, New York, New York 10016, USA. [2]Neuroscience Institute, New York University School of Medicine, New York, New York 10016, USA. [3]Department of Otolaryngology, New York University School of Medicine, New York, New York 10016, USA. [4]Department of Neuroscience and Physiology, New York University School of Medicine, New York, New York 10016, USA. [5]Department of Cell Biology, New York University School of Medicine, New York, New York 10016, USA. [6]Department of Psychiatry, New York University School of Medicine, New York, New York 10016, USA. [7]Center for Neural Science, New York University, New York, New York 10003, USA.

Figure 1 | Oxytocin enables pup retrieval. a, Retrieval behaviour. **b**, Initially naive virgins retrieving at least once <12 h after co-housing (saline: 6 out of 27 animals; oxytocin (OT): 20 out of 36 animals; $P < 0.03$, Fisher's two-tailed exact test with Bonferroni correction compared to saline; optogenetic (Opto) PVN stimulation: 5 out of 7 animals, $P < 0.05$). *$P < 0.05$. Error bars denote mean ± 95% confidence intervals. **c**, Cumulative retrieval during co-housing. **d**, Retrieval rates ($P > 0.5$, analysis of variance (ANOVA)) and speed ($P > 0.1$) were similar in dams and experienced virgins. Error bars denote mean ± s.e.m. **e**, Cumulative retrieval of saline-injected ($n = 16$) or oxytocin-injected ($n = 19$) isolated virgins (2 days after testing, saline: 2 out of 16 animals retrieved, oxytocin: 4 out of 19 animals, $P > 0.6$; 6 days, saline: 2 out of 16 animals, oxytocin: 9 out of 19 animals, $P < 0.03$).

Figure 2 | Oxytocin receptor expression in female auditory cortex. a, Antibody to mouse oxytocin receptor (OXTR-2). Top, immmunoblot of HEK cells expressing oxytocin receptors (OTR) versus control (ctrl). Bottom, OXTR-2 immunoblots of cortical lysates from wild-type (WT) and knockout (KO) animals. Red, oxytocin receptor molecular mass (43 kilodaltons (kDa)). GAPDH was used as a loading control. **b**, Immunostaining in the left auditory cortex of naive virgin. DAPI, 4′,6-diamidino-2-phenylindole. **c**, No labelling in oxytocin receptor knockouts. **d**, Left auditory cortex of eGFP–OXTR virgin co-stained for eGFP. Arrows denote double-labelled cells. **e**, Cortical interneurons express oxytocin receptors. Virgin left auditory cortex layer 5 co-stained for parvalbumin (PV) and somatostatin (SST). **f**, Co-labelled OXTR-2+ and PV+/SST+ auditory cortical cells. **g**, Left, right auditory cortex from same naive virgin. **h**, Oxytocin receptors expressed more in left auditory cortex (mothers, left: 17.4 ± 2.0%, right: 12.7 ± 2.4%, left/right asymmetry: 37.0%, $P < 0.03$, Student's two-tailed paired t-test, $n = 7$; virgins, left: 19.5 ± 1.2%, right: 14.3 ± 1.4%, asymmetry: 36.4%, $P < 0.02$, $n = 12$). Error bars denote mean ± s.e.m. Scale bars, 150 μm (**b**, **c**), 50 μm (**d**, **e**) and 100 μm (**g**).

oxytocin and oxytocin receptors[2,3,31]. Cells were unlabelled in oxytocin receptor knockout animals[32] (Fig. 2c and Extended Data Fig. 3). We also examined expression patterns in bacterial artificial chromosome (BAC) transgenic oxytocin receptor (OXTR)–enhanced green fluorescent protein (eGFP) mice[16,33] using antibodies to GFP (Fig. 2d). Around 30–40% of parvalbumin-positive and somatostatin-positive inhibitory interneurons expressed oxytocin receptors (Fig. 2e, f), suggesting that oxytocin is important for controlling cortical inhibition[15,16]. We also observed yellow fluorescent protein (YFP)-positive PVN axons in the auditory cortex of Oxt-IRES-Cre mice after viral injection, demonstrating that hypothalamic oxytocin neurons project to cortex (Extended Data Fig. 4a–c).

Notably, receptor expression in the female auditory cortex was lateralized (Fig. 2g). Significantly more cells expressed oxytocin receptors in the left auditory cortex than the right auditory cortex in mothers and naive virgins (Fig. 2h). The left auditory cortex might therefore be especially sensitive to oxytocin modulation and specialized for processing social stimuli such as pup calls. As axonal projections from PVN into cortex were not obviously lateralized (Extended Data Fig. 4d), it is likely that this anatomical specialization emerges within the cortex.

Retrieval requires the left auditory cortex

We next asked whether the left auditory cortex was functionally important for pup retrieval. We implanted cannulas into the left or

right primary auditory cortex (AI) of inexperienced virgins or dams. First, we unilaterally infused the GABA agonist muscimol to inactivate the left or right auditory cortex transiently before behavioural testing. Muscimol in the left but not right auditory cortex impaired retrieval (Fig. 3a, b and Supplementary Video 4), showing that activity in the left auditory cortex is required for retrieval. These data extend an earlier study using unilateral ear plugs to show that maternal animals have right ear/left brain advantage for recognizing pup call sounds[19].

We wondered whether oxytocin within virgin left auditory cortex might accelerate expression of retrieval. Although it is unlikely that activation of a single brain area is entirely sufficient for retrieval, sensitization of the left AI might be crucial for recognizing the significance of distress calls. Animals receiving oxytocin or optogenetically stimulated in the left auditory cortex began retrieving earlier than saline-infused animals (Fig. 3c and Supplementary Videos 5 and 6). Thus, the left auditory cortex is a major component of oxytocin-sensitive circuitry for maternal behaviour.

We then asked whether oxytocin receptor activation itself was required for behavioural performance in experienced animals. We used antagonists of oxytocin receptors (OTA or L-368,899) infused into the left auditory cortex before testing retrieval in experienced
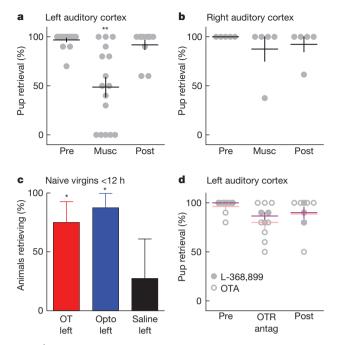
Figure 3 | Oxytocin receptors in left auditory cortex are initially required for retrieval. a, Muscimol infused into the left AI reduced retrieval by experienced animals (pre-muscimol retrieval: 96.9 ± 2.0%, 16 out of 16 animals retrieved at least once; muscimol: 48.8 ± 10.0%, $P < 0.002$, Fisher's one-way exact test, 11 out of 16 animals retrieved, $P < 0.025$). **$P < 0.01$. Error bars denote mean ± s.e.m. **b**, Muscimol in the right AI did not impair retrieval (pre-muscimol: 100.0 ± 0.0%, muscimol: 87.5 ± 12.5%, $P > 0.9$; 5 out of 5 animals retrieved). **c**, Oxytocin in the left AI of naive virgins accelerated time to first retrieval <12 h of co-housing (oxytocin: 12 out of 16 animals, $P < 0.05$; optogenetic stimulation: 7 out of 8 animals, $P < 0.04$; saline: 3 out of 11 animals). Error bars denote mean ± 95% confidence interval. **d**, Retrieval of experienced animals with the oxytocin receptor antagonists (antag) OTA (baseline: 96.3 ± 2.6%, OTA: 80.0 ± 6.8%, $P > 0.05$; 8 out of 8 animals) or L-368,899 (baseline: 100.0 ± 0.0%, L-368,899: 86.7 ± 3.3%, $P > 0.1$, $n = 3$; 3 out of 3 animals) infused into the left AI.

animals. Neither antagonist affected performance (Fig. 3d). These data suggest that oxytocin receptors might be required only when animals first begin to retrieve, but are unnecessary for expression of retrieval behaviour thereafter (analogous to requirement of NMDA receptors for long-term potentiation induction but not maintenance). Thus, after experience with pup calls during heightened cortical oxytocin levels, changes are induced in the left auditory cortex to produce enduring memory traces for maternal behaviour.

Responses to pup calls in cortical neurons

We then asked what circuit modifications in the auditory cortex are enabled by oxytocin. Our goal was to first characterize pup call responses in single neurons from maternal animals, before determining synaptic mechanisms by which oxytocin affects the virgin brain. We used *in vivo* whole-cell recordings[34–37] to measure AI pup call responses in isoflurane-anaesthetized dams, naive virgins, and experienced virgins. In 21 current-clamp and 37 cell-attached recordings, pup calls evoked stronger responses in the left AI of mothers and experienced virgins than in naive virgins (Fig. 4a–c). As responses to pure tones were comparable across groups (Fig. 4d), differences in responsiveness between experienced and inexperienced females are specific for pup calls, not simply owing to more auditory-responsive neurons in maternal cortex. Responses in experienced females were lateralized to the left AI (Extended Data Fig. 5a, b). Notably, calls evoked precise spikes in maternal animals but not naive virgins (Fig. 4e) or the right AI (Extended Data Fig. 5a, b). We quantified temporal similarity by computing trial-to-trial cross-correlation for
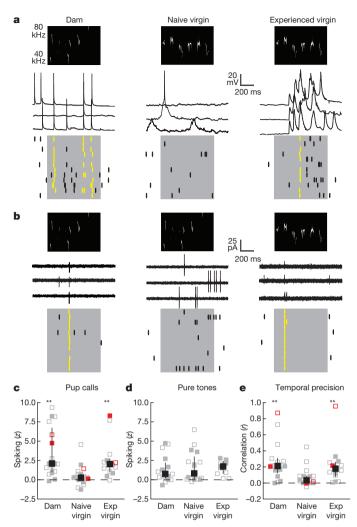
Figure 4 | Pup calls evoke reliable spikes in experienced female left AI. a, Current-clamp recordings in the left AI of dam (spiking z-score: 4.7; trial-by-trial average correlation r: 0.21), naive virgin (z-score: 0.1; r: −0.04), experienced virgin (z-score: 8.3; r: 0.27). **b**, Cell-attached recordings from dam (z-score: 5.8; r: 0.87), naive virgin (z-score: 1.4; r: 0.04), experienced virgin (z-score: 2.1; r: 0.96). **c**, Current-clamp (filled) or cell-attached (open) call-evoked responses in dams (black, z-score: 2.1 ± 1.0 (median ± s.e.m.), $n = 17$, $P < 0.0004$, $U = 288$, Wilcoxon–Mann–Whitney two-sample rank test with Bonferroni correction), naive virgins (z-score: 0.2 ± 0.3, $n = 20$), experienced virgins (z-score: 2.0 ± 0.8, $n = 14$, $P < 0.006$, $U = 240$). Error bars denote median and interquartile range. **d**, Tone-evoked responses in the left AI of dams (z-score: 0.7 ± 0.5, $n = 17$, $P > 0.7$, $U = 154$), naive virgins (z-score: 0.8 ± 0.6, $n = 17$) and experienced virgins (z-score: 1.7 ± 0.4, $n = 11$, $P > 0.6$, $U = 105$). **e**, Trial-by-trial correlation in dams (r: 0.21 ± 0.07, $n = 17$, $P < 0.002$, $U = 276$), naive virgins (r: 0.03 ± 0.03, $n = 20$) and experienced virgins (r: 0.18 ± 0.08, $n = 14$, $P < 0.012$, $U = 219$). Red symbols denote example cells shown in **a** and **b**.

spiking responses, finding higher correlations in left AI neurons from experienced females (Fig. 4e and Extended Data Fig. 5b).

To examine synaptic responses to pup calls, we made 58 *in vivo* voltage-clamp recordings from AI neurons. Substantial excitatory and inhibitory postsynaptic currents (EPSCs and IPSCs) were evoked by calls in all animals. Amplitudes of call-evoked synaptic responses (Fig. 5a, b), tone-evoked responses (Fig. 5c) and spontaneous activity (Extended Data Fig. 6a–d) were comparable across groups, suggesting that reliable call-evoked spiking in maternal left AI could not be explained simply by these neurons receiving stronger excitatory inputs.

Instead, excitation and inhibition were balanced (co-tuned and precisely timed) in the left AI neurons of experienced females but
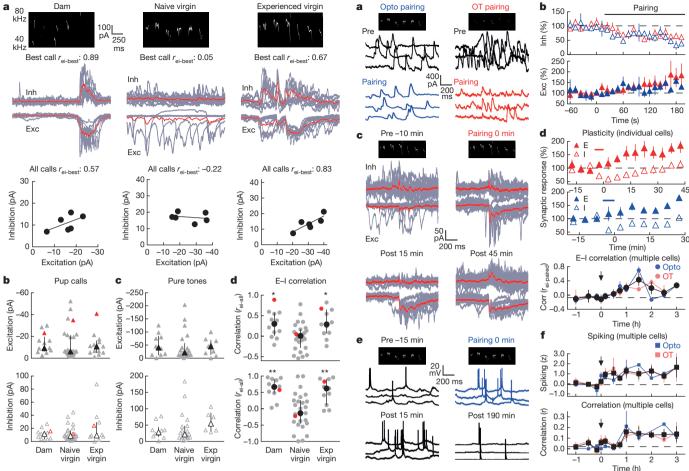
Figure 5 | Pup calls evoke correlated patterns of excitatory and inhibitory responses in left AI of experienced females. a, Voltage-clamp recordings from dam (top, best call responses, $r_{ei\text{-}best}$: 0.89; bottom, IPSCs and EPSCs across calls, $r_{ei\text{-}all}$: 0.57), naive virgin ($r_{ei\text{-}best}$: 0.05; $r_{ei\text{-}all}$: −0.22), experienced virgin ($r_{ei\text{-}best}$: 0.67; $r_{ei\text{-}all}$: 0.83). **b,** Synaptic call-evoked responses from dams (excitation: −9.3 ± 3.1 pA (median ± s.e.m.), $n = 13$, $P > 0.3$, $U = 229$, Wilcoxon–Mann–Whitney two-sample rank test with Bonferroni correction; inhibition: 11.6 ± 2.8 pA, $P > 0.5$, $U = 182$), naive virgins (excitation: −6.2 ± 3.3 pA, $n = 28$; inhibition: 8.7 ± 2.7 pA), experienced virgins (excitation: −10.8 ± 3.6 pA, $n = 13$, $P > 0.6$, $U = 205$; inhibition: 9.2 ± 9.5 pA, $P > 0.4$, $U = 171$). Red denotes cells in **a**. Error bars denote median and interquartile range. **c,** Tone-evoked responses in dams (excitation: −40.6 ± 16.7 pA, $n = 10$, $P > 0.4$, $U = 123$; inhibition: 28.7 ± 9.6 pA, $P > 0.7$, $U = 108$), naive virgins (excitation: −21.6 ± 14.8 pA, $n = 21$; inhibition: 21.8 ± 13.2 pA), experienced virgins (excitation: −45.2 ± 13.8 pA, $n = 9$, $P > 0.5$, $U = 107$; inhibition: 54.7 ± 16.4 pA, $P > 0.1$, $U = 92$). **d,** Excitatory–inhibitory (E–I) correlation of best call responses (top, $r_{ei\text{-}best}$) and across all calls (bottom, $r_{ei\text{-}all}$) dams ($r_{ei\text{-}best}$: 0.30 ± 0.12, $n = 12$, $P < 0.03$, $U = 245$; $r_{ei\text{-}all}$: 0.67 ± 0.11, $P < 0.0004$, $U = 278$), naive virgins ($r_{ei\text{-}best}$: 0.00 ± 0.08, $n = 27$; $r_{ei\text{-}all}$: −0.13 ± 0.13), experienced virgins ($r_{ei\text{-}best}$: 0.29 ± 0.13, $n = 12$, $P < 0.03$, $U = 224$; $r_{ei\text{-}all}$: 0.62 ± 0.14, $P < 0.006$, $U = 236$).

not naive virgins or the right AI. We quantified the degree of excitatory–inhibitory balance of call-evoked responses three ways: trial-by-trial similarity in patterns of excitatory or inhibitory responses (r_e, r_i), fine-scale correlation of temporal structure between EPSCs and IPSCs from best calls ($r_{ei\text{-}best}$), and overall correlation between EPSC and IPSC amplitudes across all calls ($r_{ei\text{-}all}$). First, we examined synaptic responses to best calls, and found that in experienced animals, patterns of EPSCs and IPSCs were similar and more reliable from trial-to-trial (Extended Data Fig. 6b, e).

For fine-scale excitatory–inhibitory balance, we observed that temporal profiles of call-evoked excitation and inhibition were almost identical in maternal animals. Although we could not simultaneously

Figure 6 | Oxytocin pairing modifies excitatory–inhibitory balance. a, Call-evoked IPSCs from virgin left AI neuron before and during optogenetic (blue) or oxytocin (red) pairing. **b,** Oxytocin reduced inhibition within 40–60 s (top; optogenetic pairing, $n = 4$, $P < 0.002$, Student's paired two-tailed t-test; oxytocin pairing, $n = 12$, $P < 0.04$). **c,** Voltage-clamp recording from virgin left AI neuron (pre-pairing IPSCs: 8.3 ± 1.1 pA, pre-pairing EPSCs: −8.0 ± 1.2 pA, $r_{ei\text{-}paired}$: 0.13; pairing IPSCs: 6.5 ± 1.5 pA, EPSCs: −10.4 ± 3.0 pA, $r_{ei\text{-}paired}$: −0.12; 10–15 min after pairing IPSCs: 4.9 ± 0.9 pA, $P < 0.0009$, EPSCs: −15.4 ± 1.3 pA, $P < 0.005$, $r_{ei\text{-}paired}$: −0.14; 45–50 min after pairing IPSCs: 9.6 ± 2.1 pA, $P > 0.5$, EPSCs: −13.4 ± 2.4 pA, $P < 0.002$, $r_{ei\text{-}paired}$: 0.27). Error bars denote mean ± s.e.m. **d,** Synaptic modifications. Top, individual neurons after oxytocin (EPSC increase: 43.5 ± 15.7%, $n = 10$, $P < 0.03$; IPSC decrease: −33.7 ± 7.8%, $P < 0.004$) or optogenetic pairing (EPSC increase: 47.5 ± 13.2%, $n = 6$, $P < 0.02$; IPSC decrease: −20.0 ± 4.3%, $P < 0.02$). Bottom, excitatory–inhibitory correlation ($n = 28$ cells, 17 animals; $r_{ei\text{-}paired}$ pre-pairing: −0.07 ± 0.05, $r_{ei\text{-}paired}$ 0–45 min after pairing: 0.02 ± 0.04, $P > 0.1$; $r_{ei\text{-}paired}$ 1–3 h post-pairing: 0.24 ± 0.06, $P < 0.0002$). **e,** Two current-clamp recordings from same virgin; first cell before optogenetic pairing (z-score: 0.04, r: 0.01), during pairing (z: 0.51, r: 0.00), 10–15 min post-pairing (z: 0.57, r: 0.03); second cell 180–190 min after pairing (z: 1.60, r: 0.11). **f,** Spiking. Top, call-evoked spiking ($n = 28$ cells, 13 animals; z-score pre-pairing: −0.13 ± 0.11, 0–45 min post-pairing: 0.91 ± 0.27, $P < 0.003$; z-score 1–3 h post-pairing: 1.21 ± 0.25, $P < 10^{-4}$). Bottom, trial-by-trial correlation (pre-pairing r: 0.01 ± 0.01; 0–45 min post-pairing: 0.05 ± 0.02, $P > 0.1$; 1–3 h post-pairing: 0.14 ± 0.03, $P < 10^{-4}$).

measure both excitation and inhibition in a given trial, we reasoned that because of similarity between excitation and inhibition alone, average responses would also be correlated. Indeed, temporal correlations of EPSCs and IPSCs evoked by best calls ($r_{ei\text{-}best}$) were higher in the left AI of experienced animals (Fig. 5d and Extended Data Fig. 5c). Finally, for overall correlation across vocalizations, each different call evoked a distinct EPSC/IPSC pattern. In the left AI neurons from

experienced but not naive animals, call-evoked EPSC and IPSC magnitudes were correlated (Fig. 5d and Extended Data Fig. 5c).

Temporal correlation of excitation and inhibition provides a mechanism for reliable and precisely timed spiking responses. As proof of principle, we simulated spiking evoked by different sets of synaptic call-evoked responses in a conductance-based integrate-and-fire model neuron[36]. We computed membrane potential and spiking responses and observed precisely timed spikes in cells from experienced but not naive animals (Extended Data Fig. 7), due to temporal mismatch between excitation and inhibition. Therefore, the patterns of EPSCs and IPSCs in experienced animals can account for reliable spiking observed in the maternal state.

Oxytocin modulation and cortical plasticity

Finally, we wondered how oxytocin sensitized neural circuits of virgin left AI to pup calls, to enable reliable spiking and successful retrieval in initially inexperienced animals. First we examined the neuromodulatory effects of oxytocin on cortical responses in vivo and in vitro. In voltage-clamp recordings, oxytocin reduced call-evoked IPSCs within seconds (Fig. 6a, b, open; Extended Data Fig. 8). By contrast, EPSCs were gradually modified over minutes (Fig. 6a, b, filled). Therefore, oxytocin rapidly disinhibits the auditory cortex much like acetylcholine[37–40], suggesting that oxytocin may regulate attention and increase the salience of social stimuli. These results corroborate recent findings in hippocampal slices on the effects of oxytocin[41] and oestrogen[42].

We then asked whether repetitive pairing of pup calls in the presence of oxytocin would persistently modify cortical pup call representations, effectively changing the virgin state into the maternal state. After recording responses to pup calls in virgin left AI, we paired calls for 3–5 min with either topical oxytocin application (oxytocin pairing) or optical stimulation of AI in Oxt-IRES-Cre animals (optogenetic pairing). A recording demonstrating oxytocin pairing is shown in Fig. 6c (individual trials in Extended Data Fig. 9a). Before pairing, calls evoked unreliable EPSCs and IPSCs. During and after pairing, IPSCs decreased while EPSCs potentiated, becoming more reliable. Forty-five minutes after pairing, however, IPSCs increased in strength and reliability, balancing the temporal profile of inhibition with excitation.

To examine the slower dynamics of inhibitory plasticity after pairing, we made multiple recordings in series after the first recording[36,37]. The correlation of average excitation and inhibition evoked by paired calls ($r_{ei\text{-}paired}$) steadily increased over an hour and was stable thereafter (Fig. 6d). As excitatory modifications and changes in reliability were maximal after 20–30 min, this increase in excitatory–inhibitory balance probably reflects gradual inhibitory potentiation, also evident in changes to inhibitory trial-by-trial correlations after pairing (Extended Data Fig. 9b).

Our simulations (Extended Data Fig. 7) indicated that this delayed balancing of excitation and inhibition has substantial consequences for call-evoked spiking. Specifically, spike timing precision should increase when IPSCs match the pattern of EPSCs. Consistent with this hypothesis, spiking responses were rapidly increased during and after pairing, but trial-by-trial correlations increased only after an hour (Fig. 6f). Thus, briefly pairing pup calls with cortical oxytocin triggers long-lasting changes in virgin AI circuitry, balancing inhibition with excitation to enhance call representation and perceptual salience within minutes to hours.

Discussion

Our results demonstrate a remarkable degree of functional lateralization in the mammalian brain and provide a molecular basis for this phenomenon. The left auditory cortex is specialized for recognizing the behavioural significance of infant distress calls and required for maternal retrieval of isolated pups. This is markedly similar to the asymmetry of speech processing in human temporal lobe[43,44], and supports earlier behavioural observations of auditory lateralization

in maternal mice[19]. We generated an antibody to the mouse oxytocin receptor, OXTR-2, and found more receptor expression in the left auditory cortex. A dedicated neural circuit, enriched for oxytocin receptors, might therefore be specialized for processing important social signals such as pup distress calls. This specialization would allow maternal animals to attend to their young and return pups to the nest rapidly and reliably. Furthermore, given the importance of olfactory signals in social behaviours including pup retrieval[13,23,45,46], it is likely that oxytocin also enhances olfaction[47] in combination with other cues (for example, pup calls) to improve parenting behaviour synergistically.

Although many aspects of mammalian maternal behaviour may be innate[48–50], recognition of the behavioural importance of pup calls depends on experience[21,23,25–28]. Several studies highlight differences between AI responses in maternal and pup-naive female mice. In particular, pup call responses have been found to be less reliable in pup-naive virgins[23,25–27]. Our results directly demonstrate how oxytocin paired with pup calls rapidly changes brain state, transforming weaker virgin responses into more robust and temporally precise maternal-like responses. The predominant effect of oxytocin is to reduce cortical inhibition within seconds, followed by longer-term modifications over hours proposed to be essential for balancing inhibition with excitation, enhancing spiking and successful maternal care. These synaptic dynamics are analogous to the imbalance of excitation and inhibition for tone-evoked responses in AI during early development when animals have had limited acoustic exposure[36]. After experience, however, tone-evoked excitation and inhibition become balanced over the first few weeks of life[34–37]. Our findings complement recent studies of neural circuits involved in social behaviour[23,48–50] by revealing how ethologically important behaviours with innate components can be quickly shaped and improved by experience. This may exemplify a general mechanism of neuromodulation for social behaviour.

1. Richard, P., Moos, F. & Freund-Mercier, M. J. Central effects of oxytocin. *Physiol. Rev.* **71,** 331–370 (1991).
2. Gimpl, G. & Fahrenholz, F. The oxytocin receptor system: structure, function, and regulation. *Physiol. Rev.* **81,** 629–683 (2001).
3. Insel, T. R. & Young, L. J. The neurobiology of attachment. *Nature Rev. Neurosci.* **2,** 129–136 (2001).
4. Insel, T. R. The challenge of translation in social neuroscience: a review of oxytocin, vasopressin, and affiliative behavior. *Neuron* **65,** 768–769 (2010).
5. Bartz, J. A., Zaki, J., Bolger, N. & Ochsner, K. N. Social effects of oxytocin in humans: Context and person matter. *Trends Cogn. Sci.* **15,** 301–309 (2011).
6. Churchland, P. S. & Winkielman, P. Modulating social behavior with oxytocin: how does it work? What does it mean? *Horm. Behav.* **61,** 392–399 (2012).
7. Pedersen, C. A., Ascher, J. A., Monroe, Y. L. & Prange, A. J. Oxytocin induces maternal behavior in virgin female rats. *Science* **216,** 648–650 (1982).
8. Winslow, J. T. & Insel, T. R. Social status in pairs of male squirrel monkeys determines the behavioral response to central oxytocin administration. *J. Neurosci.* **11,** 2032–2038 (1991).
9. Nishimori, K. *et al.* Oxytocin is required for nursing but is not essential for parturition or reproductive behavior. *Proc. Natl Acad. Sci. USA* **93,** 11699–11704 (1996).
10. Zak, P. J., Stanton, A. A. & Ahmadi, S. Oxytocin increases generosity in humans. *PLoS ONE* **2,** e1128 (2007).
11. Andari, E. *et al.* Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. *Proc. Natl Acad. Sci. USA* **107,** 4389–4394 (2010).
12. Chang, S. W. & Platt, M. L. Oxytocin and social cognition in rhesus macaques: Implications for understanding and treating human psychopathology. *Brain Res.* **1580,** 57–68 (2014).
13. Dulac, C., O'Connell, L. A. & Wu, Z. Neural control of maternal and paternal behaviors. *Science* **345,** 765–770 (2014).
14. Rilling, J. K. & Young, L. J. The biology of mammalian parenting and its effects on offspring social development. *Science* **345,** 771–776 (2014).
15. Yoshida, M. *et al.* Evidence that oxytocin exerts anxiolytic effects via oxytocin receptor expressed in serotonergic neurons in mice. *J. Neurosci.* **29,** 2259–2271 (2009).

16. Nakajima, M., Görlich, A. & Heintz, N. Oxytocin modulates female sociosexual behavior through a specific class of prefrontal cortical interneurons. *Cell* **159,** 295–305 (2014).
17. Sewell, G. D. Ultrasonic communication in rodents. *Nature* **227,** 410 (1970).
18. Noirot, E. Ultrasounds and maternal behavior in small rodents. *Dev. Psychobiol.* **5,** 371–387 (1972).
19. Ehret, G. Left hemisphere advantage in the mouse brain for recognizing ultrasonic communication calls. *Nature* **325,** 249–251 (1987).
20. Fichtel, I. & Ehret, G. Perception and recognition discriminated in the mouse auditory cortex by c-Fos staining. *Neuroreport* **10,** 2341–2345 (1999).
21. Ehret, G. Infant rodent ultrasounds – a gate to the understanding of sound communication. *Behav. Genet.* **35,** 19–29 (2005).
22. Crawley, J. N. Behavioral phenotyping strategies for mutant mice. *Neuron* **57,** 809–818 (2008).
23. Cohen, L., Rothschild, G. & Mizrahi, A. Multisensory integration of natural odors and sounds in the auditory cortex. *Neuron* **72,** 357–369 (2011).
24. Hofstetter, K. M. & Ehret, G. The auditory cortex of the mouse: connections of the ultrasonic field. *J. Comp. Neurol.* **323,** 370–386 (1992).
25. Liu, R. C., Linden, J. F. & Schreiner, C. E. Improved cortical entrainment to infant communication calls in mothers compared with virgin mice. *Eur. J. Neurosci.* **23,** 3087–3097 (2006).
26. Liu, R. C. & Schreiner, C. E. Auditory cortical detection and discrimination correlates with communicative significance. *PLoS Biol.* **5,** e173 (2007).
27. Rothschild, G., Cohen, L., Mizrahi, A. & Nelken, I. Elevated correlations in neuronal ensembles of mouse auditory cortex following parturition. *J. Neurosci.* **33,** 12851–12861 (2013).
28. Koch, M. & Ehret, G. Estradiol and parental experience, but not prolactin are necessary for ultrasound recognition and pup-retrieving in the mouse. *Physiol. Behav.* **45,** 771–776 (1989).
29. Irani, B. G. et al. Distribution and neurochemical characterization of protein kinase C-theta and -delta in the rodent hypothalamus. *Neuroscience* **170,** 1065–1079 (2010).
30. Wu, Z. et al. An obligate role of oxytocin neurons in diet induced energy expenditure. *PLoS ONE* **7,** e45167 (2012).
31. Knobloch, H. S. et al. Evoked axonal oxytocin release in the central amygdala attenuates fear response. *Neuron* **73,** 553–566 (2012).
32. Takayanagi, Y. et al. Pervasive social deficits, but normal parturition, in oxytocin receptor-deficient mice. *Proc. Natl Acad. Sci. USA* **102,** 16096–16101 (2005).
33. Gong, S. et al. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425,** 917–925 (2003).
34. Wehr, M. & Zador, A. M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426,** 442–446 (2003).
35. Tan, A. Y. & Wehr, M. Balanced tone-evoked synaptic excitation and inhibition in mouse auditory cortex. *Neuroscience* **163,** 1302–1315 (2009).
36. Dorrn, A. L., Yuan, K., Barker, A. J., Schreiner, C. E. & Froemke, R. C. Developmental sensory experience balances cortical excitation and inhibition. *Nature* **465,** 932–936 (2010).
37. Froemke, R. C., Merzenich, M. M. & Schreiner, C. E. A synaptic memory trace for cortical receptive field plasticity. *Nature* **450,** 425–429 (2007).
38. Kruglikov, I. & Rudy, B. Perisomatic GABA release and thalamocortical integration onto neocortical excitatory cells are regulated by neuromodulators. *Neuron* **58,** 911–924 (2008).
39. Letzkus, J. J. et al. A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* **480,** 331–335 (2011).
40. Froemke, R. C. et al. Long-term modification of cortical synapses improves sensory perception. *Nature Neurosci.* **16,** 79–88 (2013).
41. Owen, S. F. et al. Oxytocin enhances hippocampal spike transmission by modulating fast-spiking interneurons. *Nature* **500,** 458–462 (2013).
42. Rudick, C. N. & Woolley, C. S. Estrogen regulates functional inhibition of hippocampal CA1 pyramidal cells in the adult female rat. *J. Neurosci.* **21,** 6532–6543 (2001).
43. Loring, D. W. et al. Cerebral language lateralization: evidence from intracarotid amobarbital testing. *Neuropsychologia* **28,** 831–838 (1990).
44. Bishop, D. V. Cerebral asymmetry and language development: cause, correlate, or consequence? *Science* **340,** 1230531 (2013).
45. Yoon, H., Enquist, L. W. & Dulac, C. Olfactory inputs to hypothalamic neurons controlling reproduction and fertility. *Cell* **123,** 669–682 (2005).
46. Fraser, E. J. & Shah, N. M. Complex chemosensory control of female reproductive behaviors. *PLoS ONE* **9,** e90368 (2014).
47. Wacker, D. W. & Ludwig, M. Vasopressin, oxytocin, and social odor recognition. *Horm. Behav.* **61,** 259–265 (2012).
48. Lin, D. et al. Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470,** 221–226 (2011).
49. Bennur, S., Tsunada, J., Cohen, Y. E. & Liu, R. C. Understanding the neurophysiological basis of auditory abilities for social communication: a perspective on the value of ethological paradigms. *Hear. Res.* **305,** 3–9 (2013).
50. Wu, Z., Autry, A. E., Bergan, J. F., Watabe-Uchida, M. & Dulac, C. G. Galanin neurons in the medial preoptic area govern parental behavior. *Nature* **509,** 325–330 (2014).

METHODS

Behaviour. All procedures were approved under New York University Institutional Animal Care and Use Committee protocols. For measuring pup retrieval, we used 2–4-month-old C57BL/6 mothers or pup-naive virgin female mice. Dams were initially pre-screened to ensure they retrieved pups; ~1% of dams did not retrieve pups and these animals were not used for co-housing. Naive virgins were initially pre-screened for retrieval or pup mauling before co-housing; <30% of naive virgins retrieved at least one pup or mauled pups during pre-screening and these animals were excluded from subsequent behavioural studies.

Each session of testing consisted of a baseline set of 10 trials and a post-injection/infusion set of 10 trials. For baseline testing, animals were placed in a behavioural arena ($38 \times 30 \times 15$ cm) containing nesting material. Animals were given at least 20 min to acclimate before each testing session began. Three to six pups ranging from postnatal days 1 to 4 were grouped in a corner of the arena and covered with nesting material. One pup was removed from the nest and placed in an opposite corner of the arena. The experimental female was given ten trials (2 min per trial) to retrieve the displaced pup and return it back to the nest; if the displaced pup was not retrieved within 2 min, the pup was returned to the nest and the trial was scored as a failure. If the pup was successfully retrieved, the time to retrieval was scored. Another pup was then taken out of the nest, placed in an opposite corner, and the next trial was begun. After ten trials, pups were placed back into their home cage with their dam. We used an ultrasonic microphone (Avisoft) to verify that isolated pups vocalized during testing.

For post-injection testing (Fig. 1b–e), virgin females were injected intraperitoneally with oxytocin (20–50 μM in saline, 0.3 ml) or saline (0.3 ml). For post-infusion testing (Fig. 3c), cannulated virgin animals were infused with oxytocin (50 μM in saline, 1.5 μl at 1 μl min^{-1}) or saline (1.5 μl at 1 μl min^{-1}). For animals that were co-housed (Fig. 1b–d), the dam and litter were placed in the arena with the experimental virgin female and the experienced dam was given at least 5–10 min to re-acclimate. Wild-type naive virgins were randomly assigned to different groups (saline or oxytocin, co-housed or isolate) by an experimenter blind to results of data analysis. Dam retrieval rates were tested in the presence of the experimental virgin female. Afterwards, the dam was removed to her home cage and the retrieval rates of the virgin female were reassessed. Virgin female retrieval was then tested as described above, and afterwards the virgin female was returned to her home cage with dam and pups. Best retrieval performance in Fig. 1d was determined from the session with the highest retrieval rate and the session with the fastest retrieval time over all sessions for each animal. Isolated animal retrieval (Fig. 1e) was examined in the absence of an experienced dam. The isolated virgin females were housed separately from dams and pups, and did not come into contact with the dams. Retrieval was tested at the following time points: 1, 3, 6, 12, 18, 24, 36, 48 and 72 h. Power analysis was performed to determine sample size for statistical significance with a power of β: 0.7; these studies required at least six animals, satisfied in the experiments of Fig. 1. Fisher's two-tailed exact test was used for comparing numbers of animals retrieving in each group for Fig. 1c and e as these data were binomial, and Bonferroni correction used to adjust P values for multiple comparisons. Comparisons were made relative to performance of saline-injected virgin females. ANOVA was used to compare retrieval performance in experienced animals in Fig. 1d. For testing the effects of muscimol or oxytocin receptor antagonist infusion, experienced retrievers were placed in the behavioural arena. Animals were given at least 5–10 min to re-acclimate, and baseline retrieval was performed as above. The animal was then infused with muscimol (2.5 mM in saline, 1.5 μl at 1 μl min^{-1}; Fig. 3a, b) or an oxytocin receptor antagonist[51,52]: either OTA (1 μM in saline, 1.5 μl at 1 μl min^{-1}; Fig. 3d, pink line, open circles) or L-368,899 (2.5 mM in saline, 1.5 μl at 1 μl min^{-1}; Fig. 3d, purple line, filled circles). Mice were randomly assigned to different pharmacological groups by an experimenter blind to results of data analysis. Retrieval was reassessed following the infusion. After 24 h, the experimental female was given 10 baseline trials, infused with saline (1.5 μl at 1 μl min^{-1}), and retrieval re-assessed following the infusion ('post' testing). Power analysis was performed to determine sample size for statistical significance with a power of β: 0.7; these studies required at least six animals, satisfied in the experiments of Fig. 3. Wilcoxon's signed-rank matched-pairs tests were used to non-parametrically compare retrieval performance in Fig. 3a, b, d. For comparing number of animals retrieving overall, Fisher's exact test was used (two-tailed for Fig. 3c; one-tailed for Fig. 3a, as all animals initially retrieved and the number of animals retrieving could only stay the same or decrease), comparing oxytocin/optogenetic animals to saline-infused virgin performance.

Stereotaxic viral injections were performed in *Oxt*-IRES-Cre mice[29,30]. Mice were anaesthetized with 0.7–2.5% isoflurane, placed into a stereotaxic apparatus, and a craniotomy performed over PVN (from bregma: 0.72 mm posterior, 0.12 mm lateral, 4.5 mm ventral). Injections were performed with a 5 μl Hamilton syringe and a 33-gauge needle. Cre-inducible pAAV5-EF1α-DIO-

ChETA-eYFP virus[53] was injected into PVN 0.1 nl s^{-1} for a final injection volume of 1.2–1.5 μl. An optical fibre ferrule was then implanted either in PVN (4.5 mm ventral) or in electrophysiologically identified AI (100 μm ventral). The craniotomy and implant was sealed with dental cement, and the virus was given a minimum of 2 weeks to express.

Production of oxytocin receptor antibodies. A custom peptide was synthesized based on the mouse oxytocin receptor amino acid sequence: 243-EGSDAAG GAGRAALARVSSVKLISKAKI-270 in the third intracellular loop[54]. This peptide was chosen based on a high level of antigenicity using the Thermo Scientific Antigen Profiler and lack of cross-reactivity. The resulting polyclonal antisera were tested using western blot and immunohistochemistry before being further purified using affinity chromatography.

Specificity was tested via immunohistochemistry in wild-type and oxytocin receptor knockout[32] brain sections and via western blot analysis of HEK293 cells (Invitrogen) expressing oxytocin receptors (pAAV-OXTR plasmid provided by L. J. Young) and protein extracts from mouse brain. HEK cells were tested for mycoplasma around time of use (mid-2013). Lysates were immuneprecipitated and immunoblotted with OXTR-2 to further evaluate specificity of antibody.

For immunohistochemical analysis, wild-type or oxytocin receptor knockout mice were anaesthetized via intraperitoneal injection (0.1 ml per 10 g) of a ketamine–xylazine mixture containing 15 mg ml^{-1} ketamine and 5 mg ml^{-1} xylazine in 0.9% sodium chloride solution. Mice were perfused intracardially with a solution of heparin (1,000 U ml^{-1}) and PBS to prevent clotting, followed by 40 ml per mouse of freshly prepared 4% paraformaldehyde in PBS. After cryoprotection, sections were blocked and incubated with oxytocin receptor primary antibody diluted in PBS to a concentration of 1 μg ml^{-1}. Sections were incubated for 2 days at 4 °C in a moist chamber. Sections were washed with PBS (3 × 15 min at room temperature) in a staining jar and incubated for 1–2 h at room temperature in Alexa-Fluor-conjugated secondary antibodies diluted 1:500 in PBS. Unbound secondary antibodies were washed with PBS (3 × 15 min at room temperature) and sections were incubated for 10 min at room temperature with a Hoechst solution (1:10,000 stock diluted in PBS) for nuclear staining. After a final rinse, the slides were coverslipped using fluoromount G (Southern Biotechnology Associates). The brains of wild-type and knockout animals were processed together to minimize confounding factors, and parallel sections from knockout animals served as controls. For inhibitory marker co-staining, the protocol followed the procedures described above with the addition of a parvalbumin primary antibody (mouse anti-parvalbumin, Swant, 235, 1:1,000) and a somatostatin primary antibody (rat anti-somatostatin, Millipore, MAB354, 1:1,000) in the oxytocin receptor antibody solution. The secondary antibodies used in these experiments were donkey anti-rabbit Alexa 488 (Molecular Probes; 1:500), donkey anti-rat Alexa 555 (Molecular Probes; 1:500), donkey anti-mouse Alexa 647 (Molecular Probes; 1:500). As a control, omission of primary antibody and/or pre-incubation with peptide eliminated immunofluorescent labelling.

Anatomy. OXTR–eGFP mice were created by the GENSAT initiative from a BAC clone containing eGFP upstream of the oxytocin receptor gene[16,33]. Cryopreserved embryos of this line were imported from the Mutant Mouse Regional Resource Centers (MMRRC) and rederived by the NYU transgenic mouse core facility. These mice were on an FVB/N-Swiss Webster background and were bred by mating hemizygous males with wild-type females. Wild-type littermates were used as controls for eGFP antibody staining. The animals were genotyped using a strain-specific protocol provided by the MMRRC using the following primers: *Oxtr* forward: 5′-GCCACACTTTAAAGAGCCTCAA-3′; GFP reverse: 5′-TAG CGGCTGAAGCACTGCA-3′. Note that not all cells natively expressing oxytocin receptors necessarily express the transgene, due to ectopic expression or lack of regulatory elements[55].

Slides were examined and imaged using a Carl Zeiss LSM 700 confocal microscope with four solid-state lasers (405/444, 488, 555, 639 nm) and appropriate filter sets. For imaging sections co-stained with multiple antibodies, we used short-pass 555 nm (Alexa Fluor 488), short-pass 640 nm (Alexa Fluor 555), and long-pass 640 nm (Alexa Fluor 647) photomultiplier tubes. The distribution and number of immunoreactive cells in each section were determined by taking images of wild-type and knockout sections under the same laser power output, pinhole aperture, and gain. Images of left and right auditory cortex in at least three sections per brain were collected and saved for manual counts by two independent blinded observers. Mean numbers of labelled cells were calculated and compared in Fig. 2h by Student's paired two-tailed *t*-tests as data passed Kolmogorov–Smirnov normality tests. Power analysis was performed to determine sample size for statistical significance with a power of β: 0.7; these studies required at least seven animals, satisfied in the experiments of Fig. 2h. Images of sections costained with inhibitory interneuron markers were collected for each channel and merged to evaluate colocalization with oxytocin receptors by two independent blinded observers. For axon length measurements in

Extended Data Fig. 4d, YFP-positive axon segments from left and right auditory cortex sections in *Oxt*-IRES-Cre animals were quantified with ImageJ by one blinded observer. Four sections spanning anterior to posterior auditory cortex from each animal and each cortical area were examined, and confocal images consisted of a z-stack that spanned the thickness of the section. Axon counts were averaged together across the four sections from each animal, and average counts tested for statistical differences with an unpaired two-tailed Student's t-test.

Surgical preparation. For *in vivo* electrophysiology or implanting cannulas/ferrules into auditory cortex, female mice were anaesthetized with isoflurane (0.5–2.5%). A small craniotomy was performed over left or right auditory cortex with stereotaxic coordinates (from bregma in mm: 2.9 posterior, 4.0 lateral). To ensure recordings or implants were targeted to AI, we first recorded multiunit activity with tungsten electrodes. AI was mapped with pure tones (60 dB SPL, 7–79 kHz, 50 ms, 1 ms cosine on/off ramps) delivered in pseudo-random sequence at 0.5–1 Hz. For survival surgeries, a cannula or ferrule was then implanted (0.6 mm projection, dummy 0.6 mm projection, internal 0.7 mm projection) using dental acrylic, and animals were given 3–7 days to recover before behavioural testing.

For viral injections, *Oxt*-IRES-Cre animals were bred into a C57BL/6 background. Female mice 2–4 months old were anaesthetized with isoflurane (0.5–2.5%). A craniotomy was performed over the left PVN using stereotaxic coordinates (from bregma in mm: 0.7 posterior, 0.25 lateral, 4 ventral), and pAAV5-EFIα-DIO-ChETA-eYFP (1–1.2 µl) was injected (0.1 µl min^{-1}). Animals were given at least 2 weeks to recover to allow adequate expression of the ChETA variant of channelrhodopsin-2.

Electrophysiology. *In vivo* recordings were performed in a sound-attenuating chamber. Initially, auditory cortex was mapped with multi-unit recordings using a tungsten electrode to determine the tonotopic organization of the primary field AI[24]. After locating AI, *in vivo* whole-cell recordings[34–37] were made from AI neurons with a Multiclamp 700B amplifier (Molecular Devices). For current-clamp recordings, patch pipettes (4–9 MΩ) contained (in mM): 115 K-gluconate, 20 KCl, 1.5 MgCl$_2$, 10 HEPES, 10 phosphocreatine, 2 MgATP, 0.5 NaGTP, pH 7.3. For voltage-clamp recordings, pipettes contained: 130 Cs-methanesulfonate, 1 QX-314, 4 TEA-Cl, 0.5 BAPTA, 4 MgATP, 20 phosphocreatine, 10 HEPES, pH 7.2. Whole-cell recordings from AI neurons were obtained from cells located 420–800 µm below the pial surface. Data were filtered at 5 kHz, digitized at 20 kHz, and analysed with Clampfit 10 (Molecular Devices). Resting potential of AI neurons: −65.4 ± 14.7 mV; series resistance R_s: 33.1 ± 24.4 MΩ; input resistance R_i: 186.1 ± 83.1 MΩ (mean ± s.d.). Recordings were excluded from analysis if R_s or R_i changed >30% compared to the baseline period.

Pup calls were recorded from isolate pups with an ultrasonic microphone, and a library of six calls (1-s duration, maximal intensity of 60 dB SPL) was used for measuring pup call responses. For measuring spiking responses, cell-attached recordings were first high-pass filtered at 100–200 Hz. Spikes were automatically detected in current-clamp or cell-attached recordings by threshold crossing in Clampex 10. Pure tones (50-ms duration, 3-ms cosine on/off ramps) over 4–64 kHz (0.2–1.0 octave steps) were played at 60 dB SPL. Tone-evoked responses were measured in a 50-ms window starting at tone onset and compared to spontaneous activity in the 50 ms before tone onset. Call-evoked responses were measured throughout the call duration plus 200 ms, and compared to spontaneous activity in the 500 ms before call onset, normalized by computing the z-scored call-evoked firing rate relative to the spontaneous rate: $z = (\mu_{evoked} - \mu_{spontaneous})/\sigma_{spontaneous}$. Response magnitude in Figs 4 and 5 are shown for the best frequency and best call, in which 'best' is defined as the stimulus that evoked the maximal amount of spiking or excitatory current.

For measuring trial-by-trial similarity of spike trains[56], binary spike trains were smoothed by convolution with a Gaussian filter (σ: 10 ms) and trial-by-trial cross-correlations computed (~10–20 trials per call per cell); all pairwise cross-correlations were then averaged to measure r for spiking responses in Figs 4 and 5 and Extended Data Fig. 5. In raster plots of Fig. 4 and Extended Data Fig. 7, yellow events are simply illustrative and indicate spikes that are synchronous within ~10 ms on 50%+ trials. Synaptic responses and trial-by-trial correlations were measured in a similar manner for Figs 5 and 6 and Extended Data Figs 5, 6 and 9, except that responses were measured as the instantaneous current (in pA): the current integral from call onset to 200 ms after call onset (in pA × ms) and divided by the total time (in ms). Synaptic responses were not smoothed before computing zero-lag cross-correlations between all pairs of individual trials. For examining changes in spiking responses and spiking or synaptic correlations over multiple cells in Fig. 6 and Extended Data Fig. 9, measurements were made up to one hour before pairing, and at 10, 30, 45 and 60 min after pairing when possible.

Measurements at longer time periods were made whenever subsequent whole-cell recordings were obtained, separated by at least 15 min cell^{-1} thereafter.

For Figs 4 and 5 and Extended Data Figs 5–7, statistics and error bars are reported as medians ± interquartile range for spiking and synaptic responses to pup calls. As these data did not all pass Kolmogorov–Smirnov normality tests, non-parametric Wilcoxon–Mann–Whitney two-sample rank tests were used for comparing pup call responses in experienced animals to virgin animal responses (however, we note that *P* values obtained with Student's two-tailed unpaired t-tests were similar, and each of the significant differences reported here were significant under both parametric and non-parametric statistics). Power analysis was performed to determine sample size for statistical significance with a power of β: 0.7; these studies required at least seven neurons for differences in spiking and at least four neurons for differences in synaptic correlations, satisfied in the experiments of Figs 4 and 5 and Extended Data Figs 5 and 6.

For pairing pup calls with exogenous oxytocin, baseline responses to pup calls were recorded for 5–20 min. A non-preferred pup call was then presented for 1–5 min at 0.5–1 Hz in the presence of topically applied oxytocin (50 µM). For pairing calls with endogenous oxytocin release via optogenetic stimulation in *Oxt*-IRES-Cre mice, blue light pulse trains (473 nm wavelength, 10 ms pulse width duration, 30 Hz stimulation frequency, 1 s total pulse train duration) were delivered (final output powers: 10–15 mW mm^{-2} at brain surface). Pup call presentation began at optical pulse train onset. Changes in responses in Fig. 6 and Extended Data Fig. 9 were compared by Student's paired two-tailed t-tests as data passed Kolmogorov–Smirnov normality tests. Power analysis was performed to determine sample size for statistical significance with a power of β: 0.7. Modulation experiments of Fig. 6b required at least three neurons for each group, plasticity experiments of Fig. 6d required at least nine neurons for oxytocin pairing and four neurons for optogenetic pairing, and studies of synaptic and spiking correlations of Fig. 6d, f and Extended Data Fig. 9b required at least eleven neurons; these were all satisfied in the experiments of Fig. 6 and Extended Data Fig. 9.

In vitro recordings were performed in auditory cortex slices prepared from adult C57Bl/6 wild-type or *Oxt*-IRES-Cre mice. Animals were deeply anaesthetized with a 1:1 ketamine:xylazine cocktail and decapitated. The brain was rapidly placed in ice-cold dissection buffer containing (in mM): 87 NaCl, 75 sucrose, 2 KCl, 1.25 NaH$_2$PO$_4$, 0.5 CaCl$_2$, 7 MgCl$_2$, 25 NaHCO$_3$, 1.3 ascorbic acid and 10 dextrose, bubbled with 95%/5% O$_2$/CO$_2$ (pH 7.4). Slices (300–400 µm thick) were prepared with a vibratome (Leica), placed in warm dissection buffer (33–35 °C) for <30 min, then transferred to a holding chamber containing artificial cerebrospinal fluid at room temperature (ACSF, in mM: 124 NaCl, 2.5 KCl, 1.5 MgSO$_4$, 1.25 NaH$_2$PO$_4$, 2.5 CaCl$_2$ and 26 NaHCO$_3$,). Slices were kept at room temperature (22–24 °C) for at least 30 min before use. For experiments, slices were transferred to the recording chamber and perfused (2–2.5 ml min^{-1}) with oxygenated ACSF at 33 °C. Whole-cell voltage-clamp recordings were made from layer 5 pyramidal cells with an Multiclamp 700B amplifier (Molecular Devices) using IR-DIC video microscopy (Olympus). Patch pipettes (3–8 MΩ) were filled with intracellular solution (in mM: 130 Cs-methanesulfonate, 1 QX-314, 4 TEA-Cl, 0.5 BAPTA, 4 MgATP, 20 phosphocreatine, 10 HEPES, pH 7.2). Data were filtered at 2 kHz, digitized at 10 kHz, and analysed with Clampfit 10 (Molecular Devices). Recordings were excluded from analysis if R_s or R_i changed >30% compared to baseline. Focal extracellular stimulation (0.01–1.0 ms, 5–150 µA) was applied with a bipolar glass electrode 100–150 µm from the recording electrode. Mean peak IPSCs were measured in a 5–10-ms window. Changes in IPSCs were compared by Student's unpaired two-tailed t-tests as data passed Kolmogorov–Smirnov normality tests. Power analysis was performed to determine sample size with a power of β: 0.6; these studies required at least three neurons, satisfied in the experiments of Extended Data Fig. 8d.

Simulations. The simulations in Extended Data Fig. 7 used a conductance-based integrate-and-fire model neuron similar to our previous study relating synaptic currents and spike generation[36]. For simulating the spiking patterns of each cell, we used EPSCs and IPSCs from Fig. 5b. On each trial, one 1.4-s EPSC and one 1.4-s IPSC were randomly chosen from the set of recorded responses. Excitatory and inhibitory synaptic conductances (g_e and g_i, respectively) were computed from currents as previously described[34–37], and then g_e and g_i were randomly rescaled on each trial to have peak instantaneous conductance over the range 1.0–1.7 nS. Membrane voltage was computed as: $\tau_m \frac{dV}{dt} = V_{rest} - V + g_e(t)(E_e - V) + g_i(t)(E_i - V)$, with $\tau_m = 10$ ms, resting membrane potential $V_{rest} = -70$ mV, excitatory reversal potential $E_e = 0$ mV, and inhibitory reversal potential

$E_i = -70$ mV. A spike was evoked in the postsynaptic neuron if the membrane voltage reached threshold of -40 mV, at which point the membrane potential was set to -80 mV in the next time step. Spike rates and trial-by-trial correlation were determined over 25 trials (approximately the number of trials used for measuring these values in the experiments), and 12 representative trials are displayed in Extended Data Fig. 7a–c for each simulation. Code can be obtained at: http://froemkelab.med.nyu.edu/marlin_etal_simulations.

51. Huber, D., Veinante, P. & Stoop, R. Vasopressin and oxytocin excite distinct neuronal populations in the central amygdala. *Science* **308,** 245–248 (2005).

52. Dölen, G., Darvishzadeh, A., Huang, K. W. & Malenka, R. C. Social reward requires coordinated activity of nucleus accumbens oxytocin and serotonin. *Nature* **501,** 179–184 (2013).

53. Gunaydin, L. A. *et al.* Ultrafast optogenetic control. *Nature Neurosci.* **13,** 387–392 (2010).

54. Kubota, Y. *et al.* Structure and expression of the mouse oxytocin receptor gene. *Mol. Cell. Endocrinol.* **124,** 25–32 (1996).

55. Harris, J. A. *et al.* Anatomical characterization of Cre driver mice for neural circuit mapping and manipulation. *Front. Neural Circuits* **8,** 76 (2014).

56. Fellous, J. M., Tiesinga, P. H., Thomas, P. J. & Sejnowski, T. J. Discovering spike patterns in neuronal responses. *J. Neurosci.* **24,** 2989–3001 (2004).

ARTICLE

CRISPR adaptation biases explain preference for acquisition of foreign DNA

Asaf Levy[1]*, Moran G. Goren[2]*, Ido Yosef[2], Oren Auster[2], Miriam Manor[2], Gil Amitai[1], Rotem Edgar[2], Udi Qimron[2]§ & Rotem Sorek[1]§

CRISPR–Cas (clustered, regularly interspaced short palindromic repeats coupled with CRISPR–associated proteins) is a bacterial immunity system that protects against invading phages or plasmids. In the process of CRISPR adaptation, short pieces of DNA ('spacers') are acquired from foreign elements and integrated into the CRISPR array. So far, it has remained a mystery how spacers are preferentially acquired from the foreign DNA while the self chromosome is avoided. Here we show that spacer acquisition is replication-dependent, and that DNA breaks formed at stalled replication forks promote spacer acquisition. Chromosomal hotspots of spacer acquisition were confined by Chi sites, which are sequence octamers highly enriched on the bacterial chromosome, suggesting that these sites limit spacer acquisition from self DNA. We further show that the avoidance of self is mediated by the RecBCD double-stranded DNA break repair complex. Our results suggest that, in *Escherichia coli*, acquisition of new spacers largely depends on RecBCD-mediated processing of double-stranded DNA breaks occurring primarily at replication forks, and that the preference for foreign DNA is achieved through the higher density of Chi sites on the self chromosome, in combination with the higher number of forks on the foreign DNA. This model explains the strong preference to acquire spacers both from high copy plasmids and from phages.

CRISPR–Cas is an adaptive defence system in bacteria and archaea that provides acquired immunity against phages and plasmids[1–6]. It comprises multiple *cas* genes, as well as an array of short sequences ('spacers') that are mostly derived from exogenous DNA and are interleaved by short DNA repeats. The CRISPR–Cas mode of action is divided into three main stages: adaptation (or 'acquisition'), expression and interference. In the adaptation stage, a new spacer is acquired from the foreign DNA and integrated into the CRISPR array. In the expression stage, the repeat-spacer array is transcribed and further processed into short CRISPR RNAs (crRNAs). These mature crRNAs, in turn, bind to Cas proteins and form the effector protein–RNA complex. During the interference stage, the effector complex identifies foreign nucleic acid via base pairing with the crRNA and targets it for degradation.

Numerous recent studies have characterized the molecular mechanisms governing the expression and interference stages of CRISPR activity, but the molecular details of the primary adaptation stage are still elusive. It was shown that the Cas1 and Cas2 proteins are necessary for primary spacer acquisition[7], and that they form a single active complex[8]. Several systems to study spacer acquisition in the model bacterium *E. coli* have been established[7–13]. Some of these systems only express Cas1 and Cas2 but lack the CRISPR interference machinery, so that the protospacer-contributing DNA molecule is not targeted for degradation[7,8,11–13]. Strikingly, despite the lack of selection against spacer acquisition from the self chromosome, the vast majority of spacers acquired in such interference-free systems are derived from plasmid DNA[7,8,11], suggesting an intrinsic preference for the Cas1–Cas2 complex to acquire spacers from the exogenous DNA. The mechanism by which the Cas1–Cas2 complex preferentially recognizes the foreign DNA as a source for acquisition of new spacers, while avoiding taking spacers from the self chromosome, remains a major unresolved question.

Preference for exogenous DNA

We set out to understand the mechanism governing the self/non-self discrimination of the DNA source for spacer acquisition during the adaptation stage. For this, we used a previously described experimental system that monitors spacer acquisition *in vivo* in the *E. coli* type I-E CRISPR system[7,12]. In this system, *cas1* and *cas2* are carried on a plasmid (pCas1+2) and their expression is regulated by an arabinose-inducible T7 RNA polymerase (Extended Data Fig. 1). We have previously shown that expression of Cas1–Cas2 in this system leads to spacer acquisition: that is, expansion of the chromosomally encoded CRISPR I array in *E. coli* BL21-AI[7]. Since this strain of *E. coli* harbours a CRISPR array but lacks any *cas* genes on its genome, this system is interference-free, and thus does not allow 'primed' CRISPR adaptation[9,10,14,15].

After overnight growth of an *E. coli* BL21-AI culture carrying pCas1+2, we amplified the leader-proximal end of the CRISPR I array using a forward primer on the leader and a reverse primer matching spacer 2 of the native array. The amplification product, containing both native and expanded arrays, was sequenced using low-coverage Illumina technology (MiSeq) to accurately quantify the fraction of arrays that acquired a new spacer in each experiment. In parallel, high-coverage Illumina sequencing (HiSeq) was performed on gel-separated expanded arrays, to characterize the source, location and frequency of newly acquired spacers in high resolution (Extended Data Fig. 1). Overall, over 38 million newly acquired spacers were sequenced in this study (Extended Data Tables 1–3).

In cultures overexpressing Cas1–Cas2 for 16 h, 36.92% (±1.2) of the sequenced arrays contained a new spacer. Conversely, in cultures where Cas1–Cas2 were not induced, 2.61% (±0.5) of the arrays contained a new spacer after 16 h of incubation, indicating that the leakage of Cas1–Cas2 transcription (as measured by RNA sequencing; Supplementary Table 1) still resulted in spacer acquisition in a

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. [2]Department of Clinical Microbiology and Immunology, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

significant fraction of the cells (Extended Data Table 1a). Examining the origin of new spacers showed strong preference for spacer acquisition from the plasmid, with only 22.86% ($\pm$0.46) and 1.8% ($\pm$0.03) of the spacers derived from the self chromosome in the induced and non-induced cultures, respectively (Extended Data Table 1b). Considering the size of the plasmid (4.7 kilobases (kb)) and its estimated copy number of 20–40, this represents 100- to 1,000-fold enrichment for acquisition of spacers from the plasmid, compared with what is expected by the DNA content in the cell. These results also show that lower expression of Cas1–Cas2 leads to higher specificity for exogenous DNA. Therefore, most of the analyses henceforth are based on spacers acquired in conditions in which Cas1–Cas2 are expressed but not overexpressed.

Replication–dependent adaptation
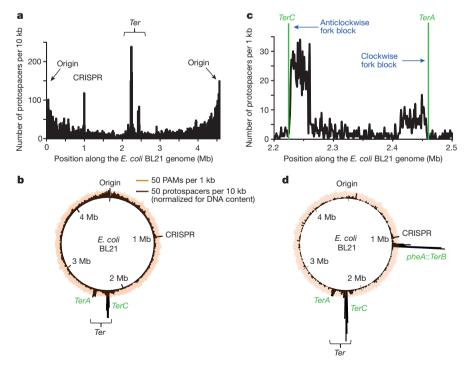
Although only a small minority of spacers was derived from the *E. coli* chromosome, the extensive number of sequenced spacers allowed us to examine chromosome-scale patterns of spacer acquisition. Remarkably, strong biases in spacer acquisition were observed, defining several protospacer hotspots (Fig. 1a). As the protospacer adjacent motif (PAM) density on the chromosome scale is largely uniform (Fig. 1b and Extended Data Fig. 2), these protospacer hotspots could not be explained by excessive localization of PAM sequences in specific areas of the genome. We further investigated each of the hotspots in search of a mechanism that would explain the observed biases.

Spacer acquisition was more pronounced at areas closer to the chromosomal origin of replication (*oriC*), with a clear gradient of reduced protospacer density as a function of the distance from *oriC* (Fig. 1a). In replicating cells, the DNA next to *oriC* is replicated first, hence the culture inevitably contains more copies of the origin-proximal DNA[16]. Indeed, upon sequencing of total genomic DNA extracted from the *E. coli* BL21-AI culture, we observed a gradient in the DNA content reminiscent of the protospacer gradient (Extended Data Fig. 2). Therefore, this *oriC*-centred spacer

acquisition bias can largely be expected based on the average DNA content in the culture and, accordingly, normalizing protospacer density to DNA content eliminated most of the *oriC*-centred protospacer gradient (Fig. 1b).

The most striking protospacer hotspot was observed around the chromosomal replication terminus (*Ter*), in two major peaks showing approximately 7- to 20-fold higher protospacer density than the surrounding area (Fig. 1b, c). The *Ter* macrodomain is the area where the two replication forks coming from opposite directions on the chromosome meet, leading to chromosome decatenation[17]. This chromosomal macrodomain contains unidirectional fork stalling sites called *Ter* sites (primarily *TerA* and *TerC*), which, during replication, stall the early-arriving replication fork until the late fork arrives from the other side[17]. We found that the primary fork-stalling sites *TerA* and *TerC* were the exact boundaries of the spacer acquisition hotspots (Fig. 1c). Moreover, the protospacer hotspots next to *Ter* sites were asymmetric relative to the fork direction of progression, with strong protospacers enrichment observed upstream of each fork stalling site and a relatively low, background protospacer density downstream of the stalled fork (Fig. 1b, c). Engineering of a native *Ter* site into the *pheA* locus on the bacterial chromosome generated a new localized protospacer hotspot, strongly supporting the idea that hotspots for spacer acquisition directly correlate with replication fork stalling sites (Fig. 1d).

The correlation between spacer acquisition biases and the replication fork stalling sites may suggest that CRISPR adaptation is promoted by active replication of the protospacer-containing DNA. We conducted a series of experiments to test this hypothesis. First, we used the replication-inhibitor quinolone nalidixic acid on *E. coli* BL21-AI cells during induction of Cas1–Cas2. As a control, we applied the RNA polymerase inhibitor rifampicin, which blocks transcription in *E. coli* but allows DNA replication (this antibiotic does not interfere with transcription of Cas1–Cas2 by the T7 RNA polymerase). Application of nalidixic acid resulted in an almost complete

Figure 1 | Chromosome-scale hotspots for spacer acquisition.
a, Distribution of protospacers across the *E. coli* BL21-AI genome. Protospacers were deduced from aligning new spacers, acquired into the CRISPR I array after 16 h growth with no arabinose, to the bacterial genome. Only unique protospacers are presented, to avoid possible biases stemming from PCR amplification of the CRISPR array. Pooled protospacers from two replicates are

presented. **b**, Protospacer density across a circular representation of the *E. coli* genome, normalized to the DNA content of the culture. Dark brown, normalized protospacer numbers; orange, PAM density. **c**, Protospacer distribution at the *Ter* region. Protospacer density is shown in 1-kb windows. **d**, Protospacer density in an *E. coli* BL21-AI in which the native 23 base pair (bp)-long *TerB* site was engineered into the *pheA* locus.
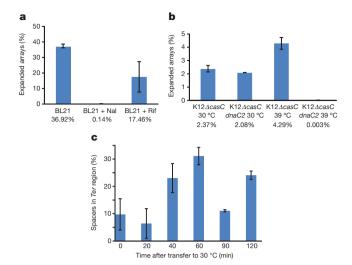
Figure 2 | Dependence of spacer acquisition on replication. a, Spacer acquisition rates in antibiotic-treated *E. coli* BL21-AI cells. Cells induced to express Cas1–Cas2 were grown for 16 h, with addition of the replication inhibitor nalidixic acid (Nal) or the transcription inhibitor rifampicin (Rif). **b**, Spacer acquisition rates of K-12*ΔcasCdnaC2* and an isogenic K-12*ΔcasC* strains during overnight Cas1–Cas2 induction. **c**, Spacer acquisition patterns measured after transfer of K-12*ΔcasCdnaC2* cells from 39 °C to 30 °C, during induction of Cas1–Cas2. For all panels, average and error margins for two biological replicates are shown.

elimination of spacer acquisition (164-fold reduction), but only an approximately twofold reduction in spacer acquisition rates was observed in the rifampicin-treated cells (Fig. 2a and Extended Data Table 1c), providing support to the hypothesis that spacer acquisition depends on DNA replication.

To substantiate these observations further, we examined the acquisition rates in *E. coli* K-12 cells carrying the temperature-sensitive allele *dnaC2* (ref. 18). In these cells, initiation of DNA replication is blocked at 39 °C but is permitted at 30 °C. These cells were transformed with a pBAD-Cas1+2 vector, in which the Cas1–Cas2 operon is directly controlled by an arabinose-inducible promoter.

Since these cells encode the full set of *cas* genes, the *casC* gene was also knocked out to avoid CRISPR interference or priming. As a control, we used an isogenic K-12 strain encoding the wild-type *dnaC* gene. After overnight growth in the replication-permissive temperature, the two strains showed similar rates of spacer acquisition. However, when the temperature-sensitive *dnaC2* cells were grown at 39 °C, acquisition was almost completely abolished, with less than 0.1% of the sequenced arrays found to be expanded (Fig. 2b and Extended Data Table 2a). These results further strengthen the hypothesis that Cas1–Cas2-mediated spacer acquisition in the *E. coli* type I-E CRISPR system requires active replication of the protospacer-containing DNA.

We next asked whether spacer acquisition preferences correlate with the position of the replication fork. For this, we transferred a culture of the temperature-sensitive *dnaC2* cells to 39 °C for 70 min. Since in this temperature replication re-initiation is inhibited, after 70 min there are no more progressing forks in these cells. We then induced Cas1–Cas2 expression for 30 min, and transferred the culture to 30 °C, resulting in synchronized initiation of replication. At these conditions, it takes the replication forks on average about 60 min to complete a full DNA replication cycle in *dnaC2* cells[19]. In accordance, we sequenced the newly acquired spacers at 20, 40, 60, 90 and 120 min following synchronous replication initiation. Strikingly, the fraction of spacers derived from the *Ter* region gradually increased with the progression of the replication cycle, reaching 31% after 60 min (compared with only 6.4% at the 20 min time point; Fig. 2c, Extended Data Fig. 3 and Extended Data Table 2b). The pattern repeated itself in the second cycle of replication (90 and 120 min; Fig. 2c). These results demonstrate temporal correlation between the predicted position of stalled replication forks and the preference to acquire spacers from that position.

Combined, the above results support a model where the Cas1–Cas2 complex has preference for acquiring spacers from the area of a stalled replication fork during DNA replication. This model is intriguing, as it largely explains the observed preference for spacer acquisition from high copy number plasmids. During DNA replication in a cell, the chromosome occupies two replication forks travelling from the *oriC* to the *Ter*, where their stalling will promote spacer acquisition. At the same replication cycle, each copy of the plasmid will occupy a
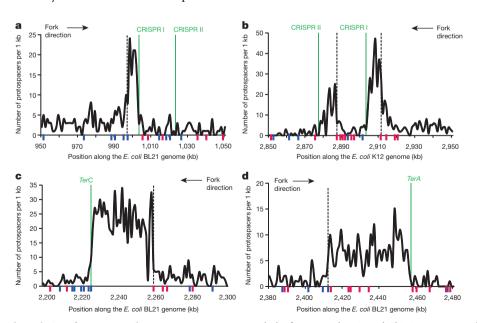
Figure 3 | Chi sites define boundaries of protospacer hotspots.
a–d, Protospacer hotspot peaks. Each panel shows a 100 kb window around a major hotspot for spacer acquisition. Short blue and red ticks mark positive- and negative-strand Chi sites, respectively. Green lines mark a replication fork stalling site (*TerA*, *TerC*) or putative stalling site (CRISPR array). Dashed lines mark the first properly oriented Chi site upstream relative to the fork stalling site. **a**, The CRISPR region in *E. coli* BL21-AI. **b**, The CRISPR region in *E. coli* K-12. **c**, The *TerC* region and **d**, the *TerA* region in *E. coli* BL21-AI. In **c**, the Chi site drawn at ~2,260 kb represents a cluster of three consecutive Chi sites found in the same 1 kb window.
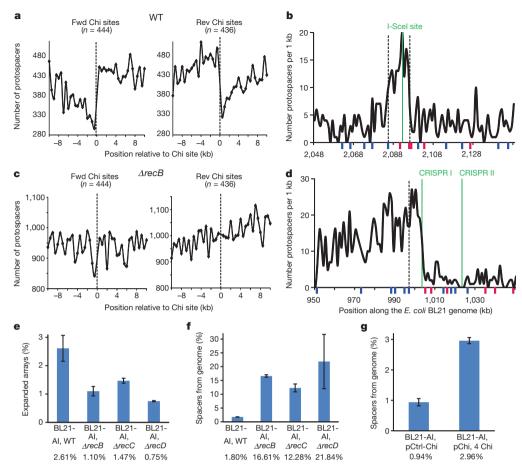
Figure 4 | Involvement of the dsDNA break repair machinery in defining spacer acquisition patterns. a, The overall number of protospacers around all Chi sites in *E. coli* BL21-AI, that are not included in the CRISPR region (950,000–1,050,000) or the *Ter* region (2–2.5 Mb), is shown in windows of 0.5 kb. WT, wild-type. **b**, Protospacer hotspot peak resulting from a dsDNA break formed by the homing endonuclease I-SceI. **c**, The overall number of protospacers around all Chi sites that are not included in the CRISPR or the *Ter* regions in a BL21-AI*ΔrecB* strain. **d**, The protospacer hotspot at the CRISPR region in the BL21-AI*ΔrecB* strain is not confined by a Chi site

(compare with the same hotspot in the wild-type strain, Fig. 3a). **e**, Adaption levels in wild-type BL21-AI and BL21-AI*ΔrecB*, *ΔrecC* or *ΔrecD* strains after overnight growth without arabinose induction of Cas1–Cas2. **f**, Percentage of new spacers derived from the self chromosome in the experiment described in **e**. **g**, Percentage of new spacers derived from the self chromosome in the presence of a plasmid containing a cluster of four Chi sites (pChi) compared with an identical plasmid lacking Chi sites (pCtrl-Chi). For **e–g**, average and error margins for two biological replicates are shown.

travelling fork, which will also be stalled during the termination of plasmid replication (in a *Ter*-independent manner[20]). Therefore, the vast majority of stalled forks in a replicating cell localize to the multiple plasmid copies, and, if spacer acquisition is promoted by fork-stalling, the probability to acquire spacers from the plasmid is much higher. The model is in line with previous observations in *Sulfolobus*, showing that spacer acquisition from an infective virus does not occur unless the viral DNA is being replicated[21].

Involvement of the DNA repair machinery

Another hotspot for spacer acquisition was observed just upstream of the CRISPR I array in the *E. coli* BL21-AI genome (Fig. 3a). This CRISPR-associated protospacer hotspot clearly depends on CRISPR activity, because no hotspot was observed near the *E. coli* BL21-AI CRISPR II array, which lacks a leader sequence and is hence inactive[7] (Fig. 3a). Indeed, in *E. coli* K-12, where both arrays are known to be active, spacer acquisition assays showed a protospacer peak upstream of each of the two arrays (Fig. 3b). The protospacer peaks at the CRISPR region resembled the peaks seen at the *Ter* sites, in the sense that they were asymmetric with respect to the replication fork direction, implying that activity at the CRISPR array forms a replication fork stalling site. Presumably the DNA nicking that occurs after the leader during insertion of a new spacer[13] stalls the replication fork,

thus generating a fork-dependent hotspot for spacer acquisition. Frequent stalling of the fork at the CRISPR would mean that the fork coming from the other direction will often be stalled for a longer time at the respective *Ter* site, *TerC*, waiting for the fork coming from the CRISPR direction to arrive (Extended Data Fig. 4). This may be one of the factors explaining why the *TerC* site is a much more pronounced protospacer hotspot than the *TerA* site (Fig. 1b, c). Another factor that can contribute to the observed *TerC*/*TerA* bias may be that the clockwise replichore in *E. coli* (*oriC* to *TerA*) is longer than the anti-clockwise one (*oriC* to *TerC*), leading the forks to naturally stall at *TerC* more often than at *TerA*.

All of the spacer acquisition hotspots described above were defined by distinct peaks of high protospacer density, with peak widths ranging between 10 and 50 kb (Fig. 3). On one end, these peaks were bounded by a fork stalling site, but the mechanism defining the boundary at the other end of the peaks was not clear. Strikingly, when searching for sequence motifs that preferentially appear at the other end of the peaks, we found that all protospacer peaks were immediately flanked by the octamer motif GCTGGTGG, which is the canonical sequence of the Chi site (Fig. 3a–d). Chi sites interact with the double-strand break repair helicase/nuclease complex RecBCD and regulate the repair activity[22]. When a double-stranded DNA (dsDNA) break occurs, RecBCD localizes to the exposed end, and then unwinds

Figure 5 | A model explaining the preference for foreign DNA in spacer acquisition. a, RecBCD localizes to a dsDNA break (DSB) and unwinds/degrades the DNA until reaching the nearest properly oriented Chi site. The RecBCD activity generates significant amounts of DNA 'debris', including short and long ssDNA fragments and degraded dsDNA, all of which may serve as substrates for spacer acquisition by Cas1–Cas2. **b**, The high density of Chi sites on the chromosome reduces spacer acquisition from self DNA. On average, the 8-bp-long Chi sites are found every 4.6 kb on the *E. coli* chromosome, 14 times more often than on random DNA. When a dsDNA break occurs on the chromosome, RecBCD DNA degradation activity will

quickly be moderated by a nearby Chi site, but a similar dsDNA break on a foreign DNA will lead to much more extensive DNA processing, providing more substrate for spacer acquisition. **c**, Preference for spacer acquisition from high-copy plasmids. In a replicating cell, most replication forks (blue circles) localize to the multiple copies of the plasmid. Since most dsDNA breaks occur during replication[23,26] at stalled replication forks[24,25], plasmid DNA would become more amenable for spacer acquisition. **d**, Most phages inject linear DNA into the infected cell. When such linear DNA is not protected, RecBCD will quickly degrade it, providing an intrinsic preference for spacer acquisition from phage DNA.

and degrades the DNA until reaching a Chi site[23]. Upon recognition of the Chi site, RecBCD generally ceases to degrade the DNA, and instead yields a single-stranded DNA that is bound by RecA and invades a homologous duplex DNA, which forms a template for completion of the missing DNA[23]. Chi sites work in an asymmetric manner, meaning that the GCTGGTGG motif will only interact with RecBCD coming from the right-end of the DNA molecule (downstream of the site), whereas the reverse complement of Chi will only interact with RecBCD complexes coming from the left-end of the DNA[22]. RecBCD indiscriminately degrades linear DNA, including phage DNA, and it was therefore suggested that this complex is one of the lines of defence against phages[23]. Since Chi sites occur every ~5 kb in the *E. coli* genome, which is about 14 times more frequent than expected by chance, these sites were suggested as markers of bacterial self, preventing RecBCD from excessively degrading the chromosome after dsDNA breaks[23].

Our results show that protospacer hotspots are defined between sites of stalled forks and Chi sites (Fig. 3). Stalled replication forks are known to be major hotspots for dsDNA breaks[24,25], and it was demonstrated that the vast majority of dsDNA breaks in bacteria occur during DNA replication[23,26]. These data therefore may imply that Cas1–Cas2 acquires spacers from degradation intermediates of RecBCD activity during the processing of dsDNA breaks that frequently occur at stalled replication forks.

Several lines of evidence support this hypothesis. First, the orientation of the Chi sites at the protospacer peaks was always consistent with the dsDNA break occurring at the fork direction rather than the other side, and the first properly oriented Chi site upstream of the stalled fork was always the site of peak boundary (Fig. 3a–d). Second, even outside the strong protospacer hotspots, there was a significant asymmetry in protospacer density upstream and downstream of Chi sites (Fig. 4a). The effect of this asymmetry was seen up to a distance of about 5–10 kb from the Chi site, consistent with an average distance of ~5 kb between Chi sites in the *E. coli* genome[22]. Third, inducing a single, site-specific dsDNA break in the chromosome using the homing endonuclease I-SceI resulted in a clear protospacer hotspot that peaked at the site of the dsDNA break and was confined by Chi sites in

the proper orientations (Fig. 4b), directly linking dsDNA breaks to spacer acquisition hotspots. Fourth, co-immunoprecipitation assays suggested that Cas1 interacts with RecB and RecC[27] (although these interactions were not verified using purified proteins), supporting a model where the Cas1–Cas2 complex is directly fed from RecBCD DNA degradation products. Finally, Cas1 was shown to efficiently bind single-stranded DNA (ssDNA), which is amply generated during RecBCD DNA processing activity[23,27].

To test whether spacer acquisition indeed depends on the activity of the RecBCD complex, we used *E. coli* strains in which *recB*, *recC* or *recD* were deleted. Deep-sequencing-based quantification of spacer acquisition rates in these mutants showed reduced acquisition in all of these deletion strains (Fig. 4e and Extended Data Table 3a). Moreover, analysis of chromosomal protospacers in these mutants showed loss of spacer acquisition asymmetry near Chi sites (Fig. 4c), resulting in broader protospacer hotspots on the self chromosome (Fig. 4d). In accordance, the fraction of spacers derived from the self chromosome was ~10-fold higher in the *recB*, *recC* and *recD* deletion strains compared with the wild-type strain (Fig. 4f and Extended Data Table 3a). These results show that CRISPR adaptation is partly dependent on the activity of the RecBCD dsDNA break repair complex, and that this activity is responsible for some of the self/non-self discrimination properties of the CRISPR adaptation process. Consistent with these results, expression of a RecBCD inhibitor protein, the product of gene 5.9 of the T7 bacteriophage[28], showed reduced spacer acquisition compared with exogenous expression of a control protein (Extended Data Fig. 5).

It is noteworthy that in *recB* and *recC* deletions, the RecBCD complex is entirely non-functional, whereas the *recD* deletion produces a complex, RecBC, that is fully functional for DNA unwinding but entirely lacks nuclease activity[23]. Our observation that the *recD* deletion mutant has poor spacer acquisition activity suggests that the nuclease activity of the RecBCD enzyme is important for spacer acquisition and implies that the degradation products generated by RecBCD during DNA processing between a dsDNA break and a Chi site may be the source of new spacers.

The involvement of Chi sites, as points where spacer acquisition activity is terminated, provides another axis for the avoidance of self

DNA in CRISPR adaptation. Since the pCas plasmid is completely devoid of Chi sites, its DNA will be fully degraded by RecBCD following any dsDNA break, providing plenty of potential substrate for Cas1–Cas2. In contrast, the high density of Chi sites on the bacterial chromosome serves for the relative avoidance of Cas1–Cas2 to acquire spacers from the chromosome, because RecBCD will only degrade the chromosomal DNA until reaching the nearest Chi site (Fig. 5a, b). Indeed, the ~10 fold higher acquisition frequency from the self chromosome seen in the *recB*, *recC* and *recD* deletion strains conforms with the natural 14-fold enrichment of Chi sites on the chromosome. To examine further whether Chi sites limit spacer acquisition, we performed spacer acquisition experiments with a plasmid that was engineered to contain a cluster of four consecutive Chi sites. As expected, an increased preference for chromosomal DNA in spacer acquisition was measured for the Chi-containing plasmid (Fig. 4g, Extended Data Table 3b and Extended Data Fig. 6).

In conclusion, these results converge to a single, unifying model that explains the preference of the CRISPR adaptation machinery to acquire spacers from foreign DNA, as well as the observed biases in spacer acquisition patterns (Fig. 5). Under this model, Cas1–Cas2 takes the DNA substrate for spacer acquisition from degradation products of RecBCD activity during the processing of dsDNA breaks. Since the vast majority of dsDNA breaks in the cell occur during DNA replication[26], with stalled replication forks being major hotspots for such breaks[24,25], high-copy-number plasmids are much more prone to spacer acquisition owing to the higher number of forks on plasmids (Fig. 5c). The high-density presence of Chi sites on the bacterial chromosome further protects it from extensive spacer acquisition (Fig. 5b). Moreover, as most phages enter the cell as a linear DNA, and since RecBCD would bind any exposed linear DNA and process it until the nearest Chi site[22], unprotected phage DNA will be a target for spacer acquisition immediately upon entry to the cell, providing an additional preference for spacer acquisition specifically from phage DNA (Fig. 5d). If entry to the cell were successful, the extensive replication activity of the phage DNA would provide another anchor for spacer acquisition from phage.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Terns, M. P. & Terns, R. M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* **14,** 321–327 (2011).
2. Westra, E. R. *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46,** 311–339 (2012).
3. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482,** 331–338 (2012).
4. Koonin, E. V. & Makarova, K. S. CRISPR-Cas: evolution of an RNA-based adaptive immune system in prokaryotes. *RNA Biol.* **10,** 679–686 (2013).
5. Sorek, R., Lawrence, C. M. & Wiedenheft, B. CRISPR-mediated adaptive immune systems in Bacteria and Archaea. *Annu. Rev. Biochem.* **82,** 237–266 (2013).
6. Barrangou, R. & Marraffini, L. A. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54,** 234–244 (2014).
7. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40,** 5569–5576 (2012).
8. Nunez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Struct. Mol. Biol.* **21,** 528–534 (2014).
9. Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7,** e35888 (2012).
10. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3,** 945 (2012).
11. Diez-Villasenor, C., Guzman, N. M., Almendros, C., Garcia-Martinez, J. & Mojica, F. J. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol.* **10,** 792–802 (2013).
12. Yosef, I. *et al.* DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proc. Natl Acad. Sci. USA* **110,** 14396–14401 (2013).
13. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Res.* **42,** 7884–7893 (2014).
14. Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol.* **10,** 716–725 (2013).
15. Fineran, P. C. *et al.* Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl Acad. Sci. USA* **111,** E1629–E1638 (2014).
16. Skovgaard, O., Bak, M., Lobner-Olesen, A. & Tommerup, N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome Res.* **21,** 1388–1393 (2011).
17. Neylon, C., Kralicek, A. V., Hill, T. M. & Dixon, N. E. Replication termination in *Escherichia coli*: structure and antihelicase activity of the Tus-Ter complex. *Microbiol. Mol. Biol. Rev.* **69,** 501–526 (2005).
18. Waldminghaus, T., Weigel, C. & Skarstad, K. Replication fork movement and methylation govern SeqA binding to the *Escherichia coli* chromosome. *Nucleic Acids Res.* **40,** 5465–5476 (2012).
19. Breier, A. M., Weier, H. U. & Cozzarelli, N. R. Independence of replisomes in *Escherichia coli* chromosomal replication. *Proc. Natl Acad. Sci. USA* **102,** 3942–3947 (2005).
20. del Solar, G. *et al.* Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* **62,** 434–464 (1998).
21. Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of *Sulfolobus*. *Mol. Microbiol.* **91,** 900–917 (2014).
22. Smith, G. R. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. *Microbiol. Mol. Biol. Rev.* **76,** 217–228 (2012).
23. Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72,** 642–671 (2008).
24. Kuzminov, A. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc. Natl Acad. Sci. USA* **98,** 8241–8246 (2001).
25. Michel, B. *et al.* Rescue of arrested replication forks by homologous recombination. *Proc. Natl Acad. Sci. USA* **98,** 8181–8188 (2001).
26. Shee, C. *et al.* Engineered proteins detect spontaneous DNA breakage in human and bacterial cells. *eLife* **2,** e01222 (2013).
27. Babu, M. *et al.* A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol. Microbiol.* **79,** 484–502 (2011).
28. Lin, L. *Study of Bacteriophage T7 Gene 5.9 and Gene 5.5*. PhD thesis, State Univ. New York (1992).

Author Contributions M.G., U.Q., A.L. and R.S. conceived and designed the research studies; M.G., I.Y., O.A., M.M., G.A. and R.E. performed the experiments; A.L., M.G., U.Q. and R.S. analysed data; A.L., M.G., U.Q. and R.S. wrote the manuscript.

Author Information RNA sequencing data are available in the National Center for Biotechnology Information Sequence Read Archive database under accession numbers SRX862155–SRX862158 in study SRP053013. Raw data of spacer sequences are accessible at http://www.weizmann.ac.il/molgen/Sorek/files/CRISPR_adaptation_2015/crispr_adaptation_2015_data.html. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to U.Q. (ehudq@post.tau.ac.il) or R.S. (rotem.sorek@weizmann.ac.il).

METHODS

No statistical methods were used to predetermine sample size.

Reagents, strains and plasmids. Luria–Bertani (LB) medium ($10\,g\,l^{-1}$ tryptone, $5\,g\,l^{-1}$ yeast extract, $5\,g\,l^{-1}$ NaCl) and agar were from Acumedia. Antibiotics and L-arabinose were from Calbiochem. Isopropyl-β-D-thiogalactopyranoside (IPTG) was from Bio-Lab. Calcium chloride ($CaCl_2$), sodium citrate (Na-citrate), restriction enzymes, T4 Polynucleotide Kinase and Phusion high fidelity DNA polymerase were from New England Biolabs. Rapid ligation kit was from Roche. Taq DNA polymerase was from LAMDA Biotech. NucleoSpin Gel and PCR Clean-up kit was from Macherey-Nagel. The bacterial strains, plasmids and oligonucleotides used in this study are listed in Supplementary Table 2.

Plasmid construction. Plasmids were constructed using standard molecular biology techniques. DNA segments were amplified by PCR. Standard digestion of the PCR products and vector by restriction enzymes was done according to the manufacturer's instructions.

pBAD plasmid encoding Cas1 and Cas2 was constructed by amplifying *cas1* and *cas2* from pWUR399 plasmid[29] using oligonucleotides IY86F and IY86R (Supplementary Table 2). The amplified DNA and pBAD18 (ref. 30) vector were both digested by SacI and SalI and ligated to yield pBAD-Cas1+2. The DNA insert was sequenced to exclude mutations introduced during cloning. pWUR plasmid encoding Cas1 and Cas2 under *lac* promoter was constructed by amplifying the *lac* promoter from pCA24N plasmid[29] using oligonucleotides SM18F and OA11R and amplifying the pCas1+2 vector using oligonucleotides IY56F and OA12F (Supplementary Table 2). The amplified products were ligated to yield pCas1+2-IPTG and sequenced to exclude mutations introduced during cloning. pWURV2 plasmid was constructed by amplifying the pCas1+2 backbone[29] using oligonucleotides IY81F and IY56R (Supplementary Table 2) followed by self-ligation. pCas1+2 plasmids harbouring four Chi sites/non-Chi sites were constructed by annealing the oligonucleotide MM1F to MM1R or MM2F to MM2R, respectively, and ligating the dsDNA product to NcoI-digested pCas1+2. pBAD33-gp5.9 plasmid encoding the T7 gene 5.9 was constructed by amplifying the 5.9 gene from the T7 bacteriophage using oligonucleotides RE45F and IY256R (Supplementary Table 2). The amplified DNA and pBAD33 (ref. 30) vector were both digested by ScaI and SalI and ligated to yield pBAD33-gp5.9.

Strain construction using recombination-based genetic engineering (recombineering). All deletion mutants used for recombineering were obtained from the Keio collection[31]. BL21-AI *recB/C/D* deletion mutants were constructed as described previously[32]. Briefly, an overnight culture of BL21-AI/pSIM6 (ref. 33) was diluted 75-fold in 250 ml LB + $100\,\mu g\,ml^{-1}$ ampicillin and aerated at 32 °C. When the attenuance, $D_{600\,nm}$, reached 0.5, the culture was heat-induced for recombination function of the prophage at 42 °C for 15 min in a shaking water bath. The induced culture was immediately cooled on ice slurry and then pelleted at $4,600g$ at 0 °C for 10 min. The pellet was washed three times in ice-cold double-distilled H_2O, then resuspended in 200 μl ice-cold double distilled H_2O and kept on ice until electroporation with ~300 ng of a gel-purified PCR product encoding the construct specified in Supplementary Table 2 containing a kanamycin-resistance cassette flanked by 50 bp homologous to the desired insertion site. A 50 μl aliquot of electrocompetent cells was used for each electroporation in a 0.2-cm cuvette at 25 μF, 2.5 kV and 200 Ω. After electroporation, the bacteria were recovered in 1 ml of 2YT ($16\,g\,l^{-1}$ bacto-tryptone, $10\,g\,l^{-1}$ yeast extract, $5\,g\,l^{-1}$ NaCl) for 2 h in a 32 °C shaking water bath and inoculated on selection plates containing $25\,\mu g\,ml^{-1}$ kanamycin. The DNA insertion into the resulting strains was confirmed by DNA sequencing of a PCR product amplifying the region.

Transductions. P1 Transductions were used for replacing *araB* with a cassette encoding the T7-RNA polymerase linked to tetracycline resistance marker, or *thr* with the *dnaC2* allele linked to Tn10 encoding tetracycline resistance marker[34], or *pheA* with the *TerB* site linked to spectinomycin[35]. P1 lysate was prepared as follows: overnight cultures of donor strain BL21-AI (for T7 RNA polymerase) or MG1655dnaC2 (for *dnaC2* allele)[34] or JJC1819 (for *pheA*::TerB-Spec)[35] were diluted 1:100 in 2.5 ml LB + 5 mM $CaCl_2$. After shaking for 1 h at 37 °C (or 30 °C for MG1655dnaC2), 0–100 μl phage P1 was added. Cultures were aerated for 1–3 h, until lysis occurred. The obtained P1 lysate was used in transduction where 100 μl fresh overnight recipient culture was mixed with 1.25 μl of 1 M $CaCl_2$ and 0–100 μl P1 phage lysate. After incubation for 30 min at 30 °C without shaking, 100 μl Na-citrate and 500 μl LB were added. Cultures were incubated at 37 °C or 30 °C for 45 or 60 min, respectively, then 3 ml of warm LB supplemented with 0.7% agar was added and the suspension was poured onto a plate containing the appropriate drug. Transductants obtained on antibiotic plates were streaked several times on selection plates and verified by PCR for the presence of the transduced DNA fragment.

Markerless insertion of I-SceI restriction site into the genome. A linear DNA containing the *Kan-sacB* cassette[36] for kanamycin resistance and sucrose

sensitivity was amplified by PCR with oligonucleotides MG53F and MG53R that provided homology to a region downstream of the *ydhQ* gene. The *Kan-sacB* cassette was inserted into DY378 strain[37] by recombineering (as described above). Colonies that were found to be resistant to kanamycin and sensitive to sucrose (that is, containing the *Kan-sacB* cassette) were picked and verified by PCR. The *Kan-sacB* cassette was transferred by P1 transduction from DY378 to BL21-AI. A second PCR was performed using oligonucleotides MG54F and MG54R that produced a short linear DNA containing the I-SceI restriction site with homology of 50 bp upstream and downstream of the *ydhQ* stop codon. Recombineering of this DNA fragment to BL21-AI, *ydhQ-Kan-sacB* resulted in kanamycin-sensitive and sucrose-resistant colonies that replaced the *Kan-sacB* cassette with the I-SceI restriction site immediately after the *ydhQ* stop codon. DNA from the resulting strain was sequence-verified for the presence of an intact I-SceI site.

CRISPR array size determination before acquisition assay. All strains underwent a preliminary validation step aimed at eliminating acquisition before induction: *E. coli* BL21-AI or K-12 harbouring pCas1+2 or pBAD-Cas1+2 plasmids, respectively, were spread on LB + $50\,\mu g\,ml^{-1}$ streptomycin or $100\,\mu g\,ml^{-1}$ ampicillin + 0.2% (w/v) glucose plates and incubated overnight at 37 °C or 30 °C (for K12ΔcasCdnaC2). A single colony was picked from each plate and used as template in a PCR amplifying CRISPR array I for BL21-AI or array II for K-12. Primers MG7R/OA1R and MG7R/MG34F were used to detect array expansion for BL21-AI and K-12, respectively (Supplementary Table 2). Only colonies that did not undergo array expansion were used in the acquisition assays described below.

Standard acquisition assay. A single colony of BL21-AI or BL21-AIΔpheA::terB, or BL21-AIΔrecB/C/D strains harbouring pCas1+2 plasmid, or BL21-AI strain harbouring pWURV2 plasmid, or K-12ΔcasC T7RNAP strain harbouring pBAD-Cas1+2 plasmid, or BL21-AI *ydhQ*-I-SceI site strain harbouring pCas1+2-IPTG and pBAD-I-SceI[38] plasmids, or BL21-AI strain harbouring pChi or pCtrl-Chi plasmids, and BL21-AI strain harbouring pCas1+2 and pBAD33-gp5.9 plasmids were inoculated in LB medium containing $50\,\mu g\,ml^{-1}$ streptomycin + 0.2% (w/v) glucose for BL21-AI strains carrying a single plasmid, or $100\,\mu g\,ml^{-1}$ ampicillin + 0.2% (w/v) glucose for K12 strain, or $100\,\mu g\,ml^{-1}$ ampicillin + $50\,\mu g\,ml^{-1}$ streptomycin + 0.2% (w/v) glucose for BL21-AI ydhQ-I-SceI site/pCas1+2-IPTG/pBAD-I-SceI strain, or $200\,\mu g\,ml^{-1}$ ampicillin + $35\,\mu g\,ml^{-1}$ chloramphenicol for BL21-AI/ pCas1+2/ pBAD33-gp5.9. Cultures were aerated at 37 °C for 16 h. Each overnight culture was diluted 1:600 in LB medium containing appropriate antibiotics with or without 0.2% (w/v) L-arabinose + 0.1 mM IPTG for pCas1+2, pChi and pCtrl-Chi harbouring strains, or 0.2% (w/v) L-arabinose for pBAD-Cas1+2 harbouring strains, or 0.02 mM IPTG and 0% L-arabinose for pCas1+2-IPTG and pBAD-I-SceI harbouring strains, or 0.4% (w/v) L-arabinose for pCas1+2 and pBAD33-gp5.9 harbouring strains. Cultures were aerated at 37 °C for an additional 16 h. DNA from these cultures was used as template (see DNA preparation for PCR) in PCRs using primers OA1F/IY130R (PCR1) and RE10RD/IY230R (PCR2) for amplifying BL21-AI CRISPR array I, or MG116F/MG34F (PCR1, see below) and RE10RD/MG115R (PCR2, see below) for amplifying K-12 array II.

Acquisition assay in the presence of antibiotics. A single colony of BL21-AI/ pCas1+2 was inoculated in LB medium containing $50\,\mu g\,ml^{-1}$ streptomycin + 0.2% (w/v) glucose and aerated at 37 °C for 16 h. The overnight cultures were diluted 1:600 in LB medium containing $50\,\mu g\,ml^{-1}$ streptomycin with or without 0.2% (w/v) L-arabinose + 0.1 mM IPTG and aerated at 37 °C. Once cultures reached a $D_{600\,nm}$ of 0.25, cells were centrifuged in a microcentrifuge for 10 min at $13,000g$ and resuspended in LB medium containing $50\,\mu g\,ml^{-1}$ streptomycin or $50\,\mu g\,ml^{-1}$ nalidixic acid or $100\,\mu g\,ml^{-1}$ rifampicin with or without 0.2% (w/v) L-arabinose + 0.1 mM IPTG. Cultures were aerated for 16 h at 37 °C, lysed and served as template for PCRs using primers OA1F/IY130R (PCR1) and RE10RD/IY230R (PCR2) for amplifying BL21-AI CRISPR array I.

Acquisition assay in replication-deficient strains. A single colony of K-12ΔcasC (control) or K-12ΔcasCdnaC2 harbouring pBAD-Cas1+2 was inoculated in LB medium containing $100\,\mu g\,ml^{-1}$ ampicillin + 0.2% (w/v) glucose and aerated at 30 °C, for 16 h. The overnight cultures were diluted 1:600 in LB medium containing $100\,\mu g\,ml^{-1}$ ampicillin + 0.2% (w/v) L-arabinose and aerated at 30 °C or 39 °C for another 16 h. Cultures were then lysed and used as template in PCRs using primers MG116F/MG34F (PCR1) and RE10RD/MG115R (PCR2) for amplifying K-12 array.

Synchronized acquisition assay. A single colony of K-12ΔcasC (control) or K-12ΔcasCdnaC2 harbouring pBAD-Cas1+2 was inoculated in LB medium containing $100\,\mu g\,ml^{-1}$ ampicillin + 0.2% (w/v) glucose and aerated at 30 °C, for 16 h. The overnight cultures were diluted 1:600 in LB medium containing $100\,\mu g\,ml^{-1}$ ampicillin + 0.2% (w/v) glucose and aerated at 30 °C until $D_{600\,nm}$ reached 0.25. Cultures were then split into six tubes and transferred to non-permissive temperature (39 °C). After 70 min, induction of Cas1–Cas2

was performed: cells were centrifuged in a standard centrifuge (4,600g, 10 min), resuspended in LB medium containing 100 µg ml^{-1} ampicillin + 0.2% (w/v) L-arabinose and aerated for an additional 30 min at 39 °C. Replication was then initiated by aerating the split cultures at 30 °C for 0, 20, 40, 60, 90 and 120 min. For replication arrest, cells were lysed and used as template in PCRs using primers MG116F/MG34F (PCR1) and RE10RD/MG115R (PCR2) for amplifying K-12 array.

DNA preparation for PCR. DNA was prepared from all cultures that underwent acquisition assays. One millilitre of each culture was centrifuged in a microcentrifuge for 1 min at 13,000g and resuspended in 100 µl LB medium. The concentrated culture underwent fast freeze in liquid nitrogen, was boiled at 95 °C for 10 min and placed on ice for 5 min. The lysate was then centrifuged in a microcentrifuge for 2 min at 13,000g; the supernatant was transferred to a new tube and served as template for PCR1 (see Preparation of DNA samples for deep sequencing).

Cultures preparation for RNA sequencing. A single colony of *E. coli* BL21-AI strain harbouring pCas1+2 plasmid was inoculated in LB medium containing 50 µg ml^{-1} streptomycin + 0.2% (w/v) glucose and aerated at 37 °C for 16 h. Each overnight culture was diluted 1:600 in LB medium containing appropriated antibiotics with or without 0.2% (w/v) L-arabinose + 0.1 mM IPTG. After overnight growth, 15 ml from each culture was centrifuged in a standard centrifuge (4,600g, 10 min), the supernatant was discarded and the pellet underwent fast freeze in liquid nitrogen. Cell pellets were then thawed and incubated at 37 °C with 300 µl 2 mg ml^{-1} lysozyme (Sigma-Aldrich catalogue number L6876-1G) in Tris 10 mM EDTA 1 mM pH 8.0, and total nucleotides were extracted using the Tri-Reagent protocol, according to the manufacturer's instructions (Molecular Research Center, catalogue number TR118). TURBO DNA-free Kit was used to eliminate DNA from the sample, according to the manufacturer's instructions (Life Technologies – Ambion catalogue number AM1907). Enrichment for messenger RNA (mRNA) was accomplished by using the Ribo-Zero rRNA Removal Kits (Illumina-Epicentre catalogue number MRZB12424). The enriched mRNA sample was then further purified using Agencourt AMPure XP magnetic beads (Beckman Coulter catalogue number A63881). Purified bacterial mRNA was then used as the starting material for the preparation of cDNA libraries for next-generation sequencing using a NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB catalogue number E7420S). The NEBNext multiplex oligonucleotides for Illumina Index primer set1 (NEB catalogue number E7335S) were used as the adapters for the library.

Total DNA purification. Overnight cultures of *E. coli* BL21-AI or K-12$\Delta casC$ T7RNAP harbouring pCas1+2 or pBAD-Cas1+2 plasmid, respectively, were diluted 1:600 and aerated for 16 h at 37 °C in LB medium containing 50 µg ml^{-1} streptomycin or 100 µg ml^{-1} ampicillin + 0.2% (w/v) glucose. These overnight cultures were then diluted 1:600 in LB medium containing 50 µg ml^{-1} streptomycin or 100 µg ml^{-1} ampicillin with 0.2% (w/v) L-arabinose + 0.1 mM IPTG or without inducers and aerated at 37 °C. Once cultures reached $D_{600\,nm} = 0.5$–0.6, 3 ml were removed and used for total DNA purification using a Macherey-Nagel NucleoSpin Tissue kit. Total DNA samples were used for deep sequencing (MiSeq).

Preparation of spacer PCR products for deep sequencing. DNA from bacterial cultures that underwent various acquisition assays was amplified in two consecutive PCRs termed PCR1 and PCR2. In PCR1, the reaction contained 20 µl of Taq 2× Master Mix master mix, 1 µl of 10 µM forward and reverse primers (see Supplementary Table 2), 4 µl of bacterial lysate and 14 µl of double-distilled water. The PCR started with 3 min at 95 °C followed by 35 cycles of 20 s at 95 °C, 20 s at 55 °C and 20 s at 72 °C. The final extension step at 72 °C was performed for 5 min. Half of the PCR1 content (20 µl) was purified using the DNA clean-up kit and were used for standard library preparation procedures followed by deep sequencing (MiSeq), while the other half (20 µl) was loaded on a 2% (w/v) agarose gel and electrophoresed for 60 min at 120 V. Following gel separation, the expanded band was excised from the gel and purified using the DNA clean-up kit. One nanogram from the extracted band served as a template for the PCR2 reaction aimed at amplifying the expanded CRISPR array products. PCR2 contained 10 µl of Taq 2 × Master Mix master mix, 0.5 µl of 10 µM forward

and reverse primers (Supplementary Table 2), 1 ng of the gel-extracted DNA from PCR1 and double-distilled water up to 20 µl. PCR2 program was identical to that of PCR1. The entire PCR2 content was loaded on a 2% (w/v) agarose gel, electrophoresed, excised and purified from the gel using the same conditions as in PCR1.

Detection of protospacer identity and acquisition level. The PCR products described above were used for preparation of Illumina sequencing libraries and were sequenced using HiSeq or MiSeq machines according to the manufacturer's instructions. Several samples were multiplexed together in the same sequencing run. Demultiplexing was performed on the basis of different Illumina barcodes and on the basis of the 3 bp barcode that was part of the original PCR primer.

Reads were mapped against the *E. coli* genome and pCAS plasmid using blastn (with parameters: -e 0.0001 -F F). For strain K-12, the Refseq accession NC_000913.2 was used; for strain BL21-AI (for which the genomic sequence is unavailable), the *E. coli* BL21-Gold(DE3)pLysS AG was used (Refseq accession NC_012947.1).

New spacer insertions were called on the basis of sequence alignments of the resulting reads. For round 1 of the PCR (Extended Data Fig. 1), alignments supporting non-acquisition events were also recorded to quantify acquisition level. If the sequence read was fully mapped to the parental CRISPR locus in the leader-proximal side, a non-acquisition event was inferred. New acquisition events were inferred if the read alignment began by a substring that was mapped to the CRISPR locus ('pre-acquisition' mapping) followed by a spacer-length substring that mapped elsewhere on the genome or the plasmid. Uninformative alignments, resulting from sequencing of the leader-distal side of the PCR amplicon, were discarded. Spacer acquisition level for a sample was defined as the number of reads supporting acquisition events divided by the number of reads either supporting or rejecting spacer acquisition.

For round 2 of the PCR (enriching for expanded arrays only, Extended Data Fig. 1) we used only unambiguously mapped protospacers (for example, spacers mapped to repetitive rRNA genes were discarded). If a spacer was mapped equally well both to the genome and the pCAS plasmid, only the plasmid protospacer position was used.

For the plots of protospacer distribution and hotspots (except for the plot in Extended Data Fig. 3), protospacer positions were recorded only once (meaning that if there were multiple spacers hitting the exact same position, the position was considered only once). This procedure was done to avoid biases stemming from PCR amplification of the CRISPR array, as well as local biases stemming from differential PAM preferences[12].

Perl and R scripts were used for data analysis. Data visualization and statistical analysis used Microsoft Excel and R, including the R circular package (http://cran.r-project.org/web/packages/circular/circular.pdf) for Fig. 1 and Extended Data Fig. 4.

29. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321,** 960–964 (2008).
30. Guzman, L. M. *et al.* Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177,** 4121–4130 (1995).
31. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2,** 1–11 (2006).
32. Sharan, S. K. *et al.* Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols* **4,** 206–223 (2009).
33. Datta, S., Costantino, N. & Court, D. L. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379,** 109–115 (2006).
34. Waldminghaus, T., Weigel, C. & Skarstad, K. Replication fork movement and methylation govern SeqA binding to the *Escherichia coli* chromosome. *Nucleic Acids Res.* **40,** 5465–5476 (2012).
35. Bidnenko, V., Ehrlich, S. D. & Michel, B. Replication fork collapse at replication terminator sequences. *EMBO J.* **21,** 3898–3907 (2002).
36. Svenningsen, S. L. *et al.* On the role of Cro in lambda prophage induction. *Proc. Natl Acad. Sci. USA* **102,** 4465–4469 (2005).
37. Yu, D. *et al.* An efficient recombination system for chromosome engineering in *Escherichia coli. Proc. Natl Acad. Sci. USA* **97,** 5978–5983 (2000).
38. Tischer, B. K. *et al.* Two-step red-mediated recombination for versatile high-efficiency markerless DNA manipulation in *Escherichia coli. Biotechniques* **40,** 191–197 (2006).

ARTICLE

Structure of the TRPA1 ion channel suggests regulatory mechanisms

Candice E. Paulsen[1]*, Jean-Paul Armache[2]*, Yuan Gao[1,2], Yifan Cheng[2] & David Julius[1]

The TRPA1 ion channel (also known as the wasabi receptor) is a detector of noxious chemical agents encountered in our environment or produced endogenously during tissue injury or drug metabolism. These include a broad class of electrophiles that activate the channel through covalent protein modification. TRPA1 antagonists hold potential for treating neurogenic inflammatory conditions provoked or exacerbated by irritant exposure. Despite compelling reasons to understand TRPA1 function, structural mechanisms underlying channel regulation remain obscure. Here we use single-particle electron cryo-microscopy to determine the structure of full-length human TRPA1 to ~4 Å resolution in the presence of pharmacophores, including a potent antagonist. Several unexpected features are revealed, including an extensive coiled-coil assembly domain stabilized by polyphosphate co-factors and a highly integrated nexus that converges on an unpredicted transient receptor potential (TRP)-like allosteric domain. These findings provide new insights into the mechanisms of TRPA1 regulation, and establish a blueprint for structure-based design of analgesic and anti-inflammatory agents.

TRP ion channels have crucial roles in somatosensation by serving as sensors for thermal and chemical stimuli[1,2]. In mammals, the TRPA1 subtype (so named for its extensive amino-terminal ankyrin repeat domain) is expressed by primary afferent nociceptors, where it detects structurally diverse noxious compounds that elicit pain and neurogenic inflammation. Such activators include pungent irritants from mustard, onion and garlic, as well as volatile environmental toxins and endogenous pro-algesic agents[3–9]. TRPA1 is also activated downstream of phospholipase-C-coupled receptors and has been proposed to function as a sensor of noxious cold[3,6,10,11]. TRPA1 is associated with persistent pain, respiratory and chronic itch syndromes[12,13] and is therefore a promising target for treating these and other neurogenic inflammatory conditions. While selective TRPA1 antagonists have been developed, their sites and mechanisms of action remain unclear.

Many TRPA1 agonists are potent electrophiles that activate the channel through covalent modification of conserved cysteine or lysine residues within the cytoplasmic N terminus[14,15]. While these and other functional properties have been deduced from electrophysiological studies of TRPA1 in whole cells, channel activity is not readily retained in excised membrane patches. This 'run down' can be mitigated if membranes are excised into solutions containing polyphosphates, suggesting that obligate cytoplasmic co-factors support TRPA1 function in intact cells[16–18]. Therefore, determining the biophysical and structural basis of polyphosphate regulation is key to understanding how TRPA1 is regulated *in vivo*, and how it can be efficiently manipulated in artificial systems for more detailed functional characterization.

A transformative step in addressing these questions would be to determine the three-dimensional (3D) atomic structure of the channel. TRP channels have posed particular challenges in this regard, probably reflecting their conformationally dynamic nature and diverse intracellular elements. Electron microscopy provides a potential means to achieve this goal, although for TRPA1 this approach has until now yielded densities of relatively low resolution (~16 Å)[19]. However, advances in single-particle electron cryo-microscopy
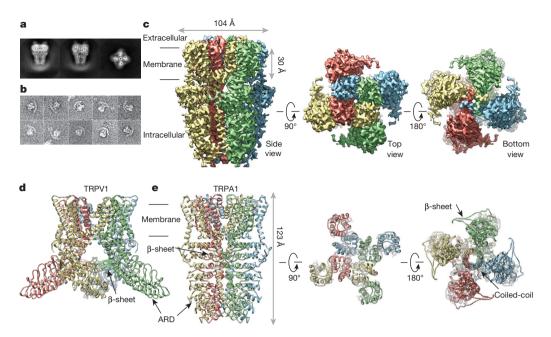
(cryo-EM) have recently enabled *de novo* structural analysis of TRPV1 to near-atomic (≤4.0 Å) resolution[20,21]. Here, we exploit this approach to determine the structure of the full-length human TRPA1 channel to ~4 Å resolution, revealing the structural basis of subunit assembly, polyphosphate action and antagonist binding.

Unique architecture of human TRPA1

A screen of TRPA1 orthologues identified the human channel as the top candidate for structural analysis based on particle size and conformational homogeneity (Extended Data Fig. 1a, b). Purified, detergent-solubilized protein appeared as homogeneous, mono-dispersed particles from which gross architectural features, such as a putative transmembrane core and extensive intracellular domain, could already be discerned (Extended Data Fig. 1b). After exchange into an amphipathic polymer, TRPA1 samples were imaged using negative-stain and single-particle cryo-EM, as previously described[21] (Fig. 1a, b and Extended Data Figs 1g and 2–5). We imaged samples under a variety of conditions, ultimately obtaining high-resolution 3D reconstructions only in the presence of an agonist (allyl isothiocyanate, AITC) or antagonists (HC-030031 with and without A-967079) to 4.24 Å, 3.9 Å and 4.7 Å, respectively, using gold-standard refinement and Fourier shell correlation (FSC) = 0.143 criterion for resolution estimation (Fig. 1c and Extended Data Figs 3–5). Indeed, even two-dimensional (2D) class averages of particles with these additives revealed elements of secondary and tertiary structure, including the tetrameric organization of the channel, well-resolved densities for the putative transmembrane core, and a central stalk flanked by convex stems that transition into a highly flexible crescent-shaped element (Fig. 1a and Extended Data Figs 3b, 4b and 5b). This latter element was routinely observed by negative-stain imaging for all orthologues examined (Fig. 1b and Extended Data Fig. 1b, g), and is thus a bona fide and conserved structural feature. In 3D reconstructions, most side chains were seen in sufficient detail to allow *de novo* atomic model building, which was initially achieved with samples treated with AITC (Fig. 1e and Extended Data Figs 3 and 6). High-resolution information suitable

[1]Department of Physiology, University of California, San Francisco, California 94158-2517, USA. [2]Keck Advanced Microscopy Laboratory, Department of Biochemistry and Biophysics, University of California, San Francisco, California 94158-2517, USA.
*These authors contributed equally to this work.

Figure 1 | 3D reconstruction of human TRPA1. a, Representative cryo-EM 2D class averages of TRPA1 (side views, left and middle; end-on view, right). **b**, Representative negative-stain particles in amphipol. **c**, 3D density map of TRPA1 from AITC-treated sample filtered to 3.5 Å resolution with each subunit colour-coded. Three views show side, top and bottom. **d**, Ribbon diagram of rat TRPV1 apo-state atomic model for comparison. **e**, Ribbon diagram of TRPA1 atomic model for residues Lys 446–Thr 1078, including the last five ankyrin repeats. Channel dimensions are indicated; side, top and bottom views are shown.

for model building spanned residues Lys 446–Thr 1078, excluding only the most distal cytoplasmic regions, a short S1–S2 linker that extends into the extracellular space, and a short helix that links a putative carboxy-terminal β-strand to the coiled-coil domain (Extended Data Fig. 7a, b). Thus, we have resolved the structure of the vast majority of the transmembrane core and proximal cytoplasmic regions involved in subunit assembly and electrophile detection. Together, these regions represent ~50% of the protein. Except where noted, discussion of the structure pertains to the AITC-treated sample.

From a bird's-eye view, TRPA1 resembles TRPV1 and distantly related voltage-gated potassium (K_v) channels: each subunit consists of six transmembrane α-helices (S1–S6) plus a re-entrant pore loop between S5 and S6; homotetramers are formed through 'domain
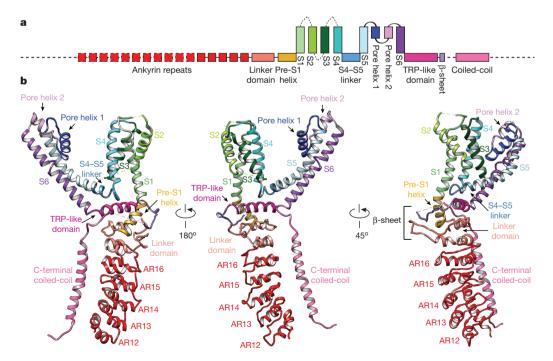
Figure 2 | Structural details of a single TRPA1 subunit. a, Linear diagram depicting major structural domains colour-coded to match ribbon diagrams below. Dashed lines and boxes denote regions for which density was insufficient to resolve detailed structure (sequence before AR12, loop containing Cys 665, S1–S2, S2–S3 and S3–S4 linkers, connection between third β-strand and coiled-coil, C terminus subsequent to coiled-coil), or where specific residues could not be definitively assigned (portion of the linker before and after the coiled-coil). **b**, Ribbon diagrams depicting three views of the TRPA1 subunit.
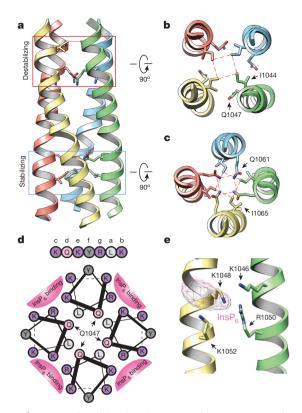
Figure 3 | C-terminal coiled-coil mediates cytosolic interactions and polyphosphate association. a, Side view of TRPA1 coiled-coil with two core glutamine residues boxed in red (destabilizing) or blue (stabilizing). **b, c**, Cross sections of coiled-coil at indicated regions with core residues depicted in stick format. Dashed red lines show residue interactions. **d**, Helical wheel presentation of residues Lys 1046–Lys 1052. Gln 1047 from each subunit is indicated with an arrow. Basic residues in 'b, e' and 'c, g' positions of neighbouring helices form the binding site for InsP$_6$. Colours differentiate class of residues: light grey, aliphatic; dark grey, aromatic; light pink, polar; purple, basic. **e**, Density for InsP$_6$ adjacent to positively-charged pocket formed by Lys 1046 and Arg 1050 from one coil, and Lys 1048 and Lys 1052 from an adjacent coil. Although not modelled, InsP$_6$ probably docks parallel to the coiled-coil such that each positively charged residue coordinates an individual phosphate moiety.

swap' interactions (Fig. 1d, e). Apart from this conserved transmembrane core, TRPA1 exhibits numerous distinctive features, particularly within the large intracellular N- and C-terminal domains, which together account for ~80% of the channel's mass (Fig. 2). For example, a C-terminal tetrameric parallel coiled-coil mediates extensive subunit interactions (Figs 1e and 2b). Also, a domain that contains five ankyrin repeats surrounds the coiled-coil and is linked with another extended feature that forms the crescent-shaped element (Figs 1e, 2b and Extended Data Fig. 8). Key cysteine residues that contribute to activation by electrophiles are located within the pre-S1 region at solvent-accessible sites, probably accounting for their relative chemical reactivity. Moreover, these residues are in close apposition to a 'TRP-like' allosteric modulatory domain characteristic of other TRP channels (but not predicted to exist in TRPA1), providing mechanistic insight into electrophile-evoked channel gating.

Coiled–coil and polyphosphate binding

A notable feature of the TRPA1 structure is a well resolved (<4 Å) tetrameric coiled-coil located at the centre of the channel, below the ion permeation pore near the C terminus, where it forms a stalk-like interaction locus for all four subunits (Fig. 1e and Extended Data Fig. 7c, d). Although coiled-coils have been shown to mediate subunit assembly for some TRP subtypes, the primary sequence within this region of TRPA1 is predicted to contain α-helices, but not a coiled-coil, per se[22]. Nonetheless, our structure reveals side-chain

interactions at 'a' and 'd' positions throughout the core, consistent with coiled-coil geometry, but distinct from the canonical coiled-coil heptad repeat, in which 'a' and 'd' positions are typically occupied by hydrophobic residues. Instead, TRPA1 contains two glutamines in staggered 'd' positions that may destabilize the structure (Gln 1047) and/or interact through intra-planar hydrogen bonds[23] (Gln 1061) (Fig. 3a–c). Furthermore, residues on the exterior surface of coiled-coils ('b, c, e–g' positions) are commonly polar or charged to facilitate inter-helix and solvent interactions, but here we see several hydrophobic or aromatic residues in these locations, of which only half interact with another channel domain (see below). This coiled-coil-based mechanism of subunit interaction differs from that seen in TRPV1, in which ankyrin repeats on one subunit form contacts with a three-stranded β-sheet on the adjacent subunit (Fig. 1d). Thus, TRP channels associate through at least two broad structural mechanisms, irrespective of whether they contain an ankyrin repeat domain (ARD).

Physiological studies have shown that soluble polyphosphates sustain TRPA1 channel activity in excised membrane patches[16–18]. Indeed, we found that inclusion of inositol hexakisphosphate (InsP$_6$) throughout channel purification was a prerequisite to obtaining mono-dispersed protein. Nevertheless, a mechanistic explanation for this phenomenon has been lacking. Remarkably, we always observed strong densities near the coiled-coil that probably correspond to InsP$_6$ (Extended Data Fig. 7d, e). Positively charged residues, namely Lys 1046 and Arg 1050 from one coil and Lys 1048 and Lys 1052 from an adjacent coil, associate with this density via four-coordinate charge–charge interactions (Fig. 3d, e). This is consistent with the observation that polyphosphates having at least four phosphate moieties are most effective at supporting TRPA1 function[16]. Interestingly, the presumptive destabilizing core residue (Gln 1047) is located between InsP$_6$-coordinating side chains (Lys 1046 and Lys 1048), and thus inter-helical cross bridging by InsP$_6$ could counteract this glutamine-mediated coiled-coil destabilization (Fig. 3a, d). This interaction is reminiscent of a previously established role for InsP$_6$ as an essential co-factor for adenosine deaminase[24], and further illustrates how cellular polyphosphates can function as primitive protein stabilization factors[25].

Pre–S1, TRP domain and reactive sites

The pre-S1 region connects the ARD to S1 and is of particular interest because it contains residues targeted by electrophilic agonists[14,15] (Fig. 4a). This region consists of two elements, including the pre-S1 helix and a preceding linker region, the primary sequence of which yields little insight into its structure or mechanistic connection to channel gating. Our 3D structures reveal an overall topology for the linker consisting of two helix-turn-helix motifs separated by two putative anti-parallel β-strands (Fig. 4a). Although TRPA1 was not predicted to contain a TRP domain because it lacks a canonical 'TRP box' motif, an α-helix directly after S6 is structurally and topologically analogous to the TRP domain in TRPV1, although located further below the inner membrane leaflet compared to TRPV1.

The linker is integrated with the overlying TRP-like domain through two main structural features: several hydrophobic interactions between stacked α-helices, and formation of a putative three-stranded β-sheet in which two strands are contributed by the linker and one by the sequence after the TRP-like domain (Fig. 4a, b). While an analogous three-stranded β-sheet in TRPV1 mediates inter-subunit interactions, this putative motif in TRPA1 seems to be a freestanding structure (Fig. 1e). Interestingly, the third β-strand connects to a short, poorly resolved α-helix that is almost buried in the inner leaflet of the membrane and forms part of a poorly resolved loop connecting the TRP-like domain with the C-terminal coiled-coil (Extended Data Fig. 7b). Many TRP channels are modulated by membrane phospholipids (such as phosphatidylinositol-4,5-bisphosphate)[26] and this connecting helix in TRPA1 may provide a structural basis for such regulation.
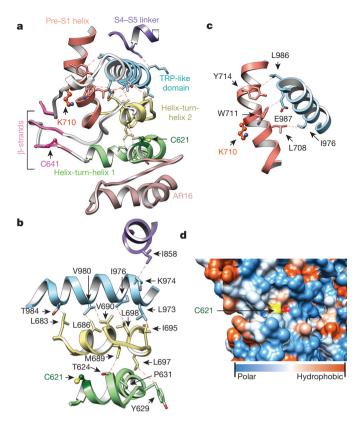
Figure 4 | Cytoplasmic domains form an integrated nexus. a, Domain architecture and web of interactions between the TRP-like domain (blue) and pre-S1 helix (orange), the overlying S4–S5 linker (purple) and underlying linker region, consisting of two helix-turn-helix motifs (green and yellow) separated by two putative anti-parallel β-strands (pink). A third β-strand (pink) is contributed by residues following the TRP-like domain. Structurally resolved reactive cysteines and lysine (Cys 621, Cys 641 and Lys 710) are shown in ball-and-stick format. The helix-turn-helices are stacked above the ARD (rose). **b**, The TRP-like domain forms hydrophobic interactions with the second helix-turn-helix motif and S4–S5 linker. The first helix-turn-helix (containing Cys 621) is integrated with the TRP-like domain through interactions with the intervening second helix-turn-helix. **c**, The TRP-like domain also interacts with the pre-S1 helix. **d**, Cys 621 is located in a closely packed pocket lined by AR16 below and the second helix-turn-helix above. Cys 621 is shown as a hydrophobicity surface.

The TRP-like domain in TRPA1 makes additional contacts with other non-contiguous structures, including the pre-S1 helix and S4–S5 linker, consistent with proposed roles for TRP domains as sites of allosteric modulation[20] (Fig. 4b, c). In TRPV1, the TRP domain interacts with the pre-S1 helix and S4–S5 linker via polar interactions[20], whereas in TRPA1 these interactions are exclusively hydrophobic, and more extensive (Fig. 4b, c). Nonetheless, our structure suggests that the TRP-like domain in TRPA1 also subserves allosteric regulation, which was not readily apparent without a 3D model.

Our structure also reveals the spatial distribution within the pre-S1 region of key cysteine and lysine residues that contribute to electrophile sensitivity: Cys 621 resides in the first helix-turn-helix; Cys 641 is located in the first strand of the putative β-sheet; Cys 665 is in a flexible loop connecting the β-strands to the second helix-turn-helix; and Lys 710 is located within the pre-S1 helix (Fig. 4a). Each of these sites is solvent accessible, and their locations within this allosteric nexus makes them well suited to detect reactive chemical agonists and transmit these events to the TRP-like domain (Fig. 4a–c). For example, Cys 621 is packed in a polar pocket between ankyrin repeat 16 (AR16) and the overlying helix-turn-helix (Fig. 4a, d). As such, electrophiles may provide a driving force for conformational change that relieves steric hindrance and/or electrostatic repulsion after

chemical modification. Furthermore, this key cysteine is located at the beginning of an α-helix and adjacent to a lysine, probably reducing the pK_a of the thiol moiety to enhance its nucleophilic character[27]. We were unable to resolve AITC-mediated adduct formation owing to insufficient resolution at these sites (estimated at 4–4.5 Å) and/or instability of the resulting modification, and thus this mechanism remains speculative until a bona fide ligand–channel complex can be clearly visualized. Moreover, we have not observed marked conformational changes within this region when comparing structures in the presence of AITC versus antagonists, but this may simply reflect lack of channel activity under conditions amenable to cryo-EM analysis. Additionally, residues in the distal N terminus (for example, Cys 421 in mouse TRPA1) or transmembrane core (for example, Ser 943 and Ile 946 in human TRPA1) have been suggested to contribute to electrophile sensitivity[15,28,29]. The former region is unresolved in our structure, and the latter residues are unlikely to be directly modified by electrophiles. The transmembrane core also contains five cysteine and lysine residues, three of which (Cys 727, Lys 771 and Cys 834) probably face the lipid environment and may therefore be modifiable by lipophilic electrophiles.

Ankyrin repeat domain

The extensive N-terminal ankyrin array is the namesake for TRPA1, yet we have little understanding of its functionality. Indeed, among vertebrate TRP proteins, TRPA1 boasts the longest ARD, variably estimated to consist of 14–18 ankyrin repeats[1,17,30,31]. In our raw micrographs (negative stain and cryo-EM) and 2D class averages, the TRPA1 N terminus is distributed into two distinct densities consisting of well-resolved convex 'stems' followed by a flexible 'crescent' (Fig. 1a, b). In all of our 3D reconstructions, we see density for five well-defined ankyrin repeats (resolved to ~4–5 Å) that contribute to the stems (Fig. 1c and Extended Data Figs 3–5). A homology model predicts that the crescent consists of 11 ankyrin repeats spanning ~100 Å, which may adopt a propeller arrangement that facilitates intersubunit interactions (Extended Data Fig. 8).

In some non-mammalian species, such as insects and snakes, TRPA1 exhibits relatively low sensitivity to electrophiles and is, instead, activated by heat[32–35]. Chimaeric and mutagenesis studies have identified regions within the ARD that specify thermal or chemical responsiveness[36,37], suggesting that the ARD can communicate with the pore. In this regard, packing interactions sterically link the ankyrin repeats in the stem with the overlying helix-turn-helix motifs of the linker region through hydrophobic and some polar interactions (Fig. 5a). This is propagated upwards and terminates within the TRP-like domain, thereby forming a network of packed interactions capable of transducing information from the ARD to the pore (Fig. 5a). Also of note are close side-chain interactions between the coiled-coil region and AR12, as well as AR16 and the first helix-turn-helix, which probably stabilize the well-resolved intracellular stems, contributing to channel assembly (Fig. 5b, c). In addition to electrophilic irritants, mammalian TRPA1 can be activated or modulated downstream of phospholipase-C-coupled receptors through increased cytosolic calcium or possibly direct interaction with Gβγ. These and other modulatory activities may also be specified by the ARD[4,6,11,18], in which case further mechanistic insights will require high-resolution structures of several TRPA1 species orthologues that include this entire domain.

Pore and antagonist binding site

The central cavity in the ion permeation pathway of AITC-treated TRPA1 (resolved to 3.5–4.5 Å) grossly resembles that of TRPV1 in exhibiting two major constrictions (Fig. 6a, b). There are, however, some notable differences. For example, the outer pore domain of TRPA1 contains two pore helices, reminiscent of bacterial voltage-gated sodium channels, where the second helix probably acts as a negatively charged conduit for attracting cations and repelling anions
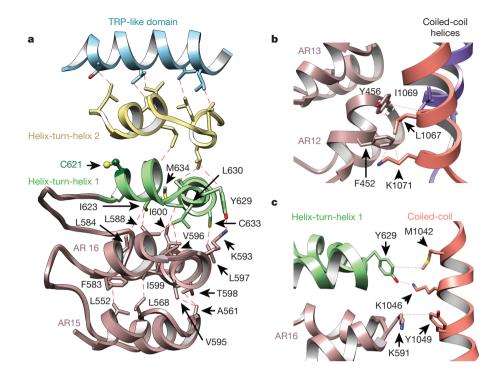
Figure 5 | Structural integration of the ARD.
a, The interdigitated convex 'stem' region of the ARD consisting of AR12–AR16 (only AR15 and AR16 are shown; rose) couples to the allosteric TRP-like domain (blue) through interactions with two intervening helix-turn-helix motifs (green and yellow) of the linker region. AR15–AR16 stacking is stabilized through hydrophobic interactions. AR16 is also connected to the overlying first helix-turn-helix motif through hydrophobic and polar interactions. **b, c**, The ARD and linker region make connections with the coiled-coil through a series of hydrophobic, polar and potentially π-cation interactions involving residues in AR12 and AR13 (**b**) as well as AR16 and the first helix-turn-helix of the linker region (**c**). Coiled-coil α-helices from the same and neighbouring subunit are coloured orange and purple, respectively.

from the channel mouth[38] (Fig. 6a). By contrast, TRPV1 and K_v channels contain only a single pore helix[21,39]. Moreover, the steeper shape of the outer pore region in TRPA1 is distinct from the wider outer vestibule seen in TRPV1 (closed state) and bacterial Na_v channels, but more reminiscent of K_v channels[38,39].

Differences are also seen in the presumptive gates. First, the upper gate in TRPV1 is formed by two residues (Gly 643 and Met 644), whereas that of TRPA1 is formed by one (Asp 915). Here, we see a restriction point measuring 7.0 Å between diagonally opposed Asp 915 residues, which is narrower than that seen in TRPV1 (7.6 Å) in the activated state, but potentially wide enough (>6 Å) to accommodate partially dehydrated calcium ions[20,40,41] (Fig. 6a, b and Extended Data Fig. 7f). Interestingly, Asp 915 has previously been implicated in controlling calcium permeability in mouse TRPA1 (ref. 18). Second, the lower gate in rat TRPV1 is formed by a single constriction in S6 at residue Ile 679, whereas that of human TRPA1 consists of two hydrophobic seals formed by Ile 957 and Val 961, creating an increasingly constricted funnel whose narrowest point (6.0 Å) is sufficient to block conduction of rehydrated cations (Fig. 6a, b and Extended Data Fig. 7g). Thus, we may have caught TRPA1 in an intermediate configuration in which the upper gate is partially open and the lower gate closed. Alternatively, the upper constriction may not constitute a regulated gate owing to a more highly scaffolded structure afforded by the second pore helix. Distinguishing between these possibilities will require stabilizing TRPA1 in distinct functional states under conditions amenable to structural analysis.

HC-030031 and A-967079 represent the two main classes of TRPA1 antagonists[7,42,43]. We determined structures of TRPA1 in the presence of HC-030031, alone, or together with A-967079. Remarkably, the double-antagonist structure revealed a unique density within a pocket formed by S5, S6 and the first pore helix (Fig. 6c and Extended Data Fig. 7h–j). This density was not observed in the structure with HC-030031 alone, and thus probably corresponds to A-967079. Phylogenetic comparisons together with molecular modelling have identified six residues required for A-967079 sensitivity[44–47], all of which surround the observed density, as does Phe 909, which is highly conserved and therefore not previously implicated in A-967079 binding (Fig. 6c). Indeed, mutation of this residue (Phe909Thr) abrogated A-967079 inhibition of

AITC-evoked responses (Fig. 6d and Extended Data Fig. 9), further corroborating assignment of this antagonist site. It has been proposed that A-967079 forms H-bond interactions with Ser 873 and/ or Thr 874, which are also located at the bottom of the putative binding pocket[47] (Fig. 6c). Consequently, orientation of the phenyl ring of the antagonist in proximity to Phe 909 may stabilize ligand binding through π–π interactions. In fact, in the double-antagonist structure, Phe 909 and additional key residues move closer to the A-967079 density, suggesting that drug binding occurs via an 'induced fit' mechanism involving movements in all three aforementioned regions that comprise the pocket (Fig. 6c, compare to AITC model in white).

A-967079 may hinder channel activity through coordinated binding to S5, S6 and first pore helix domains, which in TRPV1 are mobile elements involved in gating. As such, A-967079 may act as a molecular wedge to inhibit opening of the lower gate by impeding movement of these elements. In the case of TRPV1, classic vanilloid ligands occupy a site within the lower S2–S4 bundle. Moreover, local anaesthetics inhibit Na_v channels by binding to discrete sites along the S6 pore-lining surface to block the selectivity filter or activation gate[48,49]. Therefore, the A-967079 binding pocket described here constitutes a novel pharmacological site. Lastly, the mutations described above do not impair HC-030031 antagonism, suggesting that these two compounds bind to discrete sites (Fig. 6d and Extended Data Fig. 9g). We were unable to identify a second density corresponding to HC-030031, perhaps owing to its lower affinity, leaving its binding site and mechanism of action unresolved.

Concluding remarks

TRPA1 is a sensor for chemical irritants and a major contributor to chemonociception. We now show that key residues involved in irritant detection are solvent accessible and lie within a putative allosteric nexus converging on an unpredicted TRP-like domain, suggesting a structural basis in which TRPA1 functions as a sensitive, low-threshold electrophile receptor. An important next step is to visualize electrophile-evoked conformational changes that are associated with gating, a goal that will require robust stabilization of TRPA1 under conditions amenable to structural studies. Our analysis of how InsP_6 stabilizes the channel represents a step in this direction. The physiological purpose of
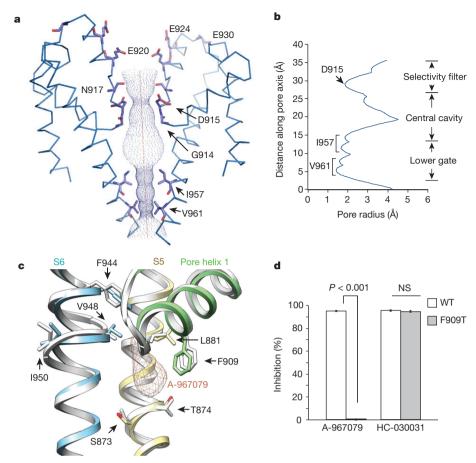
Figure 6 | The ion permeation pathway and antagonist binding site.
a, Solvent-accessible pathway along the pore of AITC-treated channel mapped with HOLE program. Asp 915 in a loop between the first and second pore helices is the sole contributor to the upper restriction, which is structurally analogous to Met 644 in TRPV1. By contrast, Gly 914 and Asn 917, structurally equivalent to Gly 643 and Asp 646 in TRPV1, do not appear to contribute to the upper constriction. A string of acidic residues in the second pore helix (Glu 920, Glu 924 and Glu 930) probably form a negatively charged conduit to attract cations and repel anions. The lower gate is formed by Ile 957 and Val 961, the former of which is analogous to Ile 679 in TRPV1. **b**, Radius of the pore as calculated through HOLE program. **c**, Cryo-EM map for the double-antagonist-treated sample contains a unique density corresponding to

A-967079 (orange) and located within a pocket formed by S5 (yellow), S6 (blue) and the first pore helix (green). Residues implicated in A-967079 antagonism are indicated, many of which line this pocket and undergo subtle conformational changes after antagonist binding (AITC model shown in white). **d**, Quantification of antagonist-mediated inhibition of AITC-evoked currents in oocytes expressing wild-type or Phe909Thr mutant TRPA1 channels. Responses were first evoked with AITC (200 μM) alone, and then in the presence of A-967079 (10 μM) or HC-030031 (100 μM). Data represent percentage of inhibition of the AITC-evoked maximal current at +80 mV ($n = 7$ independent cells per group, mean ± s.e.m., Student's t-test). Representative current traces are in Extended Data Fig. 9.

the InsP$_6$ interaction is unclear. Perhaps by stabilizing the coiled-coil domain, polyphosphates function as second messengers that, together with cytosolic calcium and G proteins, modulate TRPA1 activity when phospholipase-C-coupled receptors hydrolyse phosphatidylinositol-4,5-bisphospate to generate inositol polyphosphates.

Our study uncovers several similarities between TRPA1 and TRPV1 that probably represent pan-TRP features[20,21]. Both channels assemble as homotetramers that exhibit domain swapping within the transmembrane core, and possess an ion permeation pathway controlled by two restriction points, the lower of which involves a conserved isoleucine residue. Additional modes of intersubunit interactions are facilitated by discrete substructures within the cytoplasmic domain, although the exact nature of these contacts is protein-specific (for example, β-strand–ARD interactions, coiled-coil, crescent propeller, and so on). These cytoplasmic intersubunit interactions may regulate channel assembly and/or facilitate concerted conformational changes after co-factor binding or agonist-evoked gating, akin to domain swapping in the transmembrane core. Additionally, an α-helix subsequent to S6 (TRP domain or structural analogue) probably operates as a conserved allosteric regulatory structure that engages in

extensive interactions with pore-forming domains. High-resolution structures of additional TRP subtypes will no doubt expand on this preliminary list of common and distinct features that account for the great functional diversity of TRP ion channels.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Julius, D. TRP channels and pain. *Annu. Rev. Cell Dev. Biol.* **29**, 355–384 (2013).
2. Wang, H. & Woolf, C. J. Pain TRPs. *Neuron* **46**, 9–12 (2005).
3. Bandell, M. *et al.* Noxious cold ion channel TRPA1 is activated by pungent compounds and bradykinin. *Neuron* **41**, 849–857 (2004).
4. Bautista, D. M. *et al.* TRPA1 mediates the inflammatory actions of environmental irritants and proalgesic agents. *Cell* **124**, 1269–1282 (2006).
5. Bautista, D. M. *et al.* Pungent products from garlic activate the sensory ion channel TRPA1. *Proc. Natl Acad. Sci. USA* **102**, 12248–12252 (2005).
6. Jordt, S. E. *et al.* Mustard oils and cannabinoids excite sensory nerve fibres through the TRP channel ANKTM1. *Nature* **427**, 260–265 (2004).
7. McNamara, C. R. *et al.* TRPA1 mediates formalin-induced pain. *Proc. Natl Acad. Sci. USA* **104**, 13525–13530 (2007).

8. Taylor-Clark, T. E. *et al.* Prostaglandin-induced activation of nociceptive neurons via direct interaction with transient receptor potential A1 (TRPA1). *Mol. Pharmacol.* **73,** 274–281 (2008).

9. Trevisani, M. *et al.* 4-Hydroxynonenal, an endogenous aldehyde, causes pain and neurogenic inflammation through activation of the irritant receptor TRPA1. *Proc. Natl Acad. Sci. USA* **104,** 13519–13524 (2007).

10. Caspani, O. & Heppenstall, P. A. TRPA1 and cold transduction: an unresolved issue? *J. Gen. Physiol.* **133,** 245–249 (2009).

11. Wilson, S. R. *et al.* TRPA1 is required for histamine-independent, Mas-related G protein-coupled receptor-mediated itch. *Nature Neurosci.* **14,** 595–602 (2011).

12. Andrade, E. L., Meotti, F. C. & Calixto, J. B. TRPA1 antagonists as potential analgesic drugs. *Pharmacol. Ther.* **133,** 189–204 (2012).

13. Kremeyer, B. *et al.* A gain-of-function mutation in TRPA1 causes familial episodic pain syndrome. *Neuron* **66,** 671–680 (2010).

14. Hinman, A., Chuang, H. H., Bautista, D. M. & Julius, D. TRP channel activation by reversible covalent modification. *Proc. Natl Acad. Sci. USA* **103,** 19564–19568 (2006).

15. Macpherson, L. J. *et al.* Noxious compounds activate TRPA1 ion channels through covalent modification of cysteines. *Nature* **445,** 541–545 (2007).

16. Kim, D. & Cavanaugh, E. J. Requirement of a soluble intracellular factor for activation of transient receptor potential A1 by pungent chemicals: role of inorganic polyphosphates. *J. Neurosci.* **27,** 6500–6509 (2007).

17. Nilius, B., Prenen, J. & Owsianik, G. Irritating channels: the case of TRPA1. *J. Physiol. (Lond.)* **589,** 1543–1549 (2011).

18. Wang, Y. Y., Chang, R. B., Waters, H. N., McKemy, D. D. & Liman, E. R. The nociceptor ion channel TRPA1 is potentiated and inactivated by permeating calcium ions. *J. Biol. Chem.* **283,** 32691–32703 (2008).

19. Cvetkov, T. L., Huynh, K. W., Cohen, M. R. & Moiseenkova-Bell, V. Y. Molecular architecture and subunit organization of TRPA1 ion channel revealed by electron microscopy. *J. Biol. Chem.* **286,** 38168–38176 (2011).

20. Cao, E., Liao, M., Cheng, Y. & Julius, D. TRPV1 structures in distinct conformations reveal activation mechanisms. *Nature* **504,** 113–118 (2013).

21. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–112 (2013).

22. Samad, A. *et al.* The C-terminal basic residues contribute to the chemical- and voltage-dependent activation of TRPA1. *Biochem. J.* **433,** 197–204 (2011).

23. Woolfson, D. N. The design of coiled-coil structures and assemblies. *Adv. Protein Chem.* **70,** 79–112 (2005).

24. Macbeth, M. R. *et al.* Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. *Science* **309,** 1534–1539 (2005).

25. Gray, M. J. *et al.* Polyphosphate is a primordial chaperone. *Mol. Cell* **53,** 689–699 (2014).

26. Rohacs, T. Phosphoinositide regulation of TRP channels. *Handb. Exp. Pharmacol.* **223,** 1143–1176 (2014).

27. Paulsen, C. E. & Carroll, K. S. Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. *Chem. Rev.* **113,** 4633–4679 (2013).

28. Chen, J. *et al.* Molecular determinants of species-specific activation or blockade of TRPA1 channels. *J. Neurosci.* **28,** 5063–5071 (2008).

29. Moparthi, L. *et al.* Human TRPA1 is intrinsically cold- and chemosensitive with and without its N-terminal ankyrin repeat domain. *Proc. Natl Acad. Sci. USA* **111,** 16901–16906 (2014).

30. Jaquemar, D., Schenker, T. & Trueb, B. An ankyrin-like protein with transmembrane domains is specifically lost after oncogenic transformation of human fibroblasts. *J. Biol. Chem.* **274,** 7325–7333 (1999).

31. Zayats, V. *et al.* Regulation of the transient receptor potential channel TRPA1 by its N-terminal ankyrin repeat domain. *J. Mol. Model.* **19,** 4689–4700 (2013).

32. Gracheva, E. O. *et al.* Molecular basis of infrared detection by snakes. *Nature* **464,** 1006–1011 (2010).

33. Sokabe, T., Tsujiuchi, S., Kadowaki, T. & Tominaga, M. *Drosophila* painless is a Ca^{2+}-requiring channel activated by noxious heat. *J. Neurosci.* **28,** 9929–9938 (2008).

34. Viswanath, V. *et al.* Opposite thermosensor in fruitfly and mouse. *Nature* **423,** 822–823 (2003).

35. Zhong, L. *et al.* Thermosensory and nonthermosensory isoforms of *Drosophila melanogaster* TRPA1 reveal heat-sensor domains of a thermoTRP Channel. *Cell Rep* **1,** 43–55 (2012).

36. Cordero-Morales, J. F., Gracheva, E. O. & Julius, D. Cytoplasmic ankyrin repeats of transient receptor potential A1 (TRPA1) dictate sensitivity to thermal and chemical stimuli. *Proc. Natl Acad. Sci. USA* **108,** E1184–E1191 (2011).

37. Jabba, S. *et al.* Directionality of temperature activation in mouse TRPA1 ion channel can be inverted by single-point mutations in ankyrin repeat six. *Neuron* **82,** 1017–1031 (2014).

38. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475,** 353–358 (2011).

39. Long, S. B., Campbell, E. B. & Mackinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K$^+$ channel. *Science* **309,** 897–903 (2005).

40. Susankova, K., Ettrich, R., Vyklicky, L., Teisinger, J. & Vlachova, V. Contribution of the putative inner-pore region to the gating of the transient receptor potential vanilloid subtype 1 channel (TRPV1). *J. Neurosci.* **27,** 7578–7585 (2007).

41. Voets, T., Janssens, A., Droogmans, G. & Nilius, B. Outer pore architecture of a Ca^{2+}-selective TRP channel. *J. Biol. Chem.* **279,** 15223–15230 (2004).

42. McGaraughty, S. *et al.* TRPA1 modulation of spontaneous and mechanically evoked firing of spinal neurons in uninjured, osteoarthritic, and inflamed rats. *Mol. Pain* **6,** 14 (2010).

43. Petrus, M. *et al.* A role of TRPA1 in mechanical hyperalgesia is revealed by pharmacological inhibition. *Mol. Pain* **3,** 40 (2007).

44. Banzawa, N. *et al.* Molecular basis determining inhibition/activation of nociceptive receptor TRPA1: a single amino acid dictates species-specific actions of the most potent mammalian trpa1 antagonists. *J. Biol. Chem.* **289,** 31927–31939 (2014).

45. Klement, G. *et al.* Characterization of a ligand binding site in the human transient receptor potential ankyrin 1 pore. *Biophys. J.* **104,** 798–806 (2013).

46. Nakatsuka, K. *et al.* Identification of molecular determinants for a potent mammalian TRPA1 antagonist by utilizing species differences. *J. Mol. Neurosci.* **51,** 754–762 (2013).

47. Xiao, B. *et al.* Identification of transmembrane domain 5 as a critical molecular determinant of menthol sensitivity in mammalian TRPA1 channels. *J. Neurosci.* **28,** 9640–9651 (2008).

48. Bagnéris, C. *et al.* Prokaryotic NavMs channel as a structural and functional model for eukaryotic sodium channel antagonism. *Proc. Natl Acad. Sci. USA* **111,** 8428–8433 (2014).

49. Catterall, W. A. Structure and function of voltage-gated sodium channels at atomic resolution. *Exp. Physiol.* **99,** 35–51 (2014).

Author Contributions C.E.P. expressed and purified protein samples, determined conditions to enhance protein stability, and performed functional studies. Y.G. carried out initial negative-stain analysis and characterization of cryo-EM conditions. J.-P.A. carried out detailed cryo-EM experiments, including data acquisition and processing. C.E.P. and J.-P.A. built the atomic model on the basis of cryo-EM maps. All authors contributed to experimental design, data analysis and manuscript preparation.

METHODS

FSEC screening of TRPA1 species orthologues. Ten TRPA1 species orthologues (human, mouse, rat, fruit bat, vampire bat, *Caenorhabditis elegans*, zebrafish isoform A, zebrafish isoform B, rat snake and rattlesnake) were screened by fluorescence size exclusion chromatography (FSEC) as previously described[50] (Extended Data Fig. 1a). In brief, orthologues were subcloned into a pCDNA3.1(+) vector containing an N-terminal enhanced green fluorescent protein (eGFP) tag. Then, 16 h after transient transfection in the presence of 3 μM ruthenium red, HEK293 cells were washed with PBS, collected in buffer A (50 mM Tris, 37.5 mM sucrose, 5 mM EDTA, 1 mM β-mercaptoethanol, 1 mM phytic acid (InsP$_6$, Sigma), 1× complete protease inhibitor cocktail (Roche), pH 7.5) and lysed by sonication on ice. Cell debris was cleared by centrifugation (8,000g × 20 min) at 4 °C and membrane pellets were collected by ultracentrifugation (100,000g × 1 h) at 4 °C. The resulting pellets were resuspended in buffer B (20 mM HEPES, 150 mM NaCl, 10% glycerol, 1 mM dithiothreitol (DTT), 1 mM InsP$_6$, 1× protease inhibitor cocktail, pH 8.0) and membranes were solubilized with 10 mM MNG-3 for 1.5 h at 4 °C (ref. 51). Detergent-insoluble material was removed by ultracentrifugation as above and the supernatant was separated on a superose 6 column in buffer C (20 mM HEPES, 150 mM NaCl, 10% glycerol, 1 mM DTT, 1 mM InsP$_6$, 0.5 mM MNG-3, pH 8.0). The eluent from the superose 6 column was passed through a Jasco fluorometer (FP-2020 Plus) fitted with a flow cell, as described in the manufacturer's instructions. The fluorometer settings were: excitation 488 nm, emission 509 nm. On the basis of the FSEC screen, the vampire bat TRPA1 orthologue was initially pursued for structural studies; however, this sample did not show sufficient particle homogeneity nor did it retain the ARD diffuse skirt subsequent to exchange into amphipols. Therefore, the ten species orthologues were further screened by negative-stain imaging of maltose-binding protein (MBP)-tagged constructs to identify the human TRPA1 orthologue as the most promising target for structural analysis.

Protein expression and purification. TRPA1 species orthologues were subcloned into a pFastBac vector containing a mammalian cytomegalovirus promoter (P$_{CMV}$) and an N-terminal MBP tag for baculovirus transduction-based expression in HEK293S GnTi⁻ cells as previously described[21]. For protein expression, HEK293 GnTi⁻ cells (lacking *N*-acetylglucosaminyltransferase I), grown in suspension at 37 °C in an orbital shaker, were transduced when cell density reached 1×10^6–1.5×10^6 per ml and were supplemented with 3 μM ruthenium red. Sodium butyrate was added to the culture 24 h after transduction at a final concentration of 10 mM to boost protein expression. Cells were collected 48 h after transduction and were broken by passing through an emulsifier twice in buffer A. Cell debris was cleared by centrifugation (8,000g × 20 min). Membranes were collected by ultracentrifugation (200,000g × 1 h) and solubilized in buffer B. Membranes were stored at −80 °C or solubilized with 10 mM MNG-3 for 2 h at 4 °C. Detergent-insoluble material was removed by ultracentrifugation (30,000g × 30 min) and the supernatant was mixed with amylose resin (New England Biolabs) for 2 h at 4 °C. The resin was washed with buffer D (20 mM HEPES, 150 mM NaCl, 1 mM DTT, 1 mM InsP$_6$, pH 8.0) containing 0.5 mM MNG-3 and bound protein was eluted with the same buffer supplemented with 40 mM maltose. For orthologue screening, eluted protein was used directly for negative-stain electron microscopy EM imaging (Extended Data Fig. 1b). For preparation of samples for cryo-EM analysis, MBP-tagged TRPA1 was mixed with PMAL-C8 (Affymetrix) at a 1:3 (w/w) dilution with gentle agitation overnight at 4 °C. Detergent was removed with Bio-Beads SM-2 (Bio-Rad) for 1 h at 4 °C, and the beads were subsequently removed over a disposable polyprep column. The eluent was cleared by centrifugation before further purification on a superose 6 column in buffer E (20 mM HEPES, 150 mM NaCl, 1 mM DTT, 1 mM InsP$_6$, pH 8.0). The peak corresponding to tetrameric MBP-tagged TRPA1 was collected for analysis by cryo-EM. Protein was supplemented with 2.5% (v/v) glycerol and mixed with agonist (100 μM AITC, Sigma) or antagonists (50 μM HC-030031 and 50 μM A-967079, 2% (v/v) DMSO, Tocris) for 10 min at room temperature before applying to grids. In PMAL-C8, purified TRPA1 remained stable and monodispersed (Extended Data Figs 1g, 3a, 4a and 5a). Each subunit of native (untagged) TRPA1 consists of 1,119 residues with a mass of 127.5 kilodaltons (kDa). Chemical structures and molecular masses of important compounds used in this study are shown in Extended Data Fig. 9i.

Cell imaging and electrophysiology. Around 16 h after transient transfection, HEK293 cells were loaded with fura-2-acetoxymethylester in physiological Ringer's buffer (140 mM NaCl, 5 mM KCl, 2 mM CaCl$_2$, 2 mM MgCl$_2$, 10 mM HEPES, pH 7.4) for ratiometric calcium imaging. Activity of TRPA1 constructs was determined by application of AITC (250 μM) and sensitivity to A-967079 was initially examined by application of antagonist (10 μM) 1 min before AITC addition. The Phe909Thr mutant was generated by site-directed mutagenesis with a QuikChange Lightning kit (Agilent). Oocyte recordings were carried out as previously described[36]. Oocytes were injected with 1 ng complementary RNA encod-

ing wild-type or mutant *TRPA1* channels and currents were recorded the next day. Drugs were applied by superfusion using an AutoMate perfusion system. Currents were first evoked with AITC (200 μM) to obtain maximal response, and then the percentage inhibition was determined by subsequent co-application of AITC with antagonist (Extended Data Fig. 9c–h). Data in Fig. 6 represent the percentage inhibition of the AITC-evoked maximal current at +80 mV ($n = 7$ independent cells per group, mean ± s.e.m., Student's *t*-test).

Electron microscopy data acquisition. The electron microscopy data acquisition and processing has been performed as described[21]. Detergent solubilized MBP–TRPA1 particles were monodispersed as assessed by negative-stain electron microscopy (Extended Data Fig. 1b). Grids for negative-stain electron microscopy were prepared following the established protocol[52]. Specifically, 2.5 μl of purified TRPA1 was applied to glow-discharged electron microscopy grids covered by a thin layer of continuous carbon film and was stained with 0.75% (w/v) uranyl formate. Negatively stained electron microscopy grids were imaged on a Tecnai T12 microscope (FEI Company) operated at 120 kV at nominal magnification of ×52,000 using a 4 k × 4 k CCD camera (UltraScan 4000, Gatan), corresponding to a pixel size of 2.02 Å on the specimen.

For cryo-EM, 2.5 μl of purified TRPA1 sample at a concentration of 0.5 mg ml⁻¹ was applied to a glow discharged Quantifoil holey carbon grid (1.2 μm hole size, 400 mesh), blotted inside a Vitrobot Mark I (FEI Company) using 7 s blotting time with 100% humidity, and then plunge-frozen in liquid ethane cooled by liquid nitrogen. Cryo-EM images were collected at liquid nitrogen temperature on a Tecnai TF20 electron microscope (FEI) operated at 200 kV using a CT3500 side entry holder (Gatan), following the low-dose procedure; images were recorded at a nominal magnification of ×80,000 using a phosphor-scintillator-based TemF816 8 k × 8 k CMOS camera (TVIPS GmbH), corresponding to a pixel size of 0.9 Å per pixel on the specimen. Images were recorded with a defocus in the range from 1.8 to 3.5 μm.

Subsequently, three data sets were collected on TF30 Polara electron microscope (FEI Company) operated at 300 kV, equipped with a K2 Summit direct electron detector camera (Gatan). Images were recorded using super-resolution counting mode following an established protocol[53]. Specifically, images from TF30 were recorded at a nominal magnification of ×31,000, corresponding to a calibrated physical pixel size of 1.22 Å per pixel on the specimen. The dose rate on the camera was set to be 8.2 counts (corresponding to 9.9 electrons) per physical pixel per second. The total exposure time was 6 s, leading to a total accumulated dose of 41 electrons per Å² on the specimen. Each image was fractionated into 30 subframes, each with an accumulation time of 0.2 s per frame. All dose-fractionated cryo-EM images were recorded using a semi-automated acquisition program UCSFImage4 (written by X. Li). Images were recorded with a defocus in a range from 1.5 to 2.8 μm.

Image processing. SamViewer, an interactive image analysis program written in Python, was used for all 2D image display and particle picking. Negative-stain EM images were 2 × 2 binned for manual particle picking. Defocus was determined using CTFFIND3 (ref. 54). Individual particles were cut out and normalized to have a mean of 0 and a standard deviation of 1. For 2D classification, particles were first corrected for contrast transfer function (CTF) by flipping the phase using 'ctfapply' (written by X. Li), and subject to six cycles of correspondence analysis, *k*-means classification and multi-reference alignment, using SPIDER operations 'CA S', 'CL KM' and 'AP SH'[55].

Low-dose images of frozen hydrated MBP–TRPA1 collected on TF20 were binned 2 × 2, resulting in a pixel size of 1.89 Å for image processing. For particle picking, images were further 2 × 2 binned to a pixel size of 3.8 Å. An *ab initio* 3D reconstruction was first determined using the TF20 data set, using a probabilistic initial 3D model generation procedure (PRIME) implemented in Simple package[56]. This reconstruction served as an initial reference model for subsequent maximum likelihood-based 3D classification and auto-refinement procedure implemented in RELION[57]. The final 3D reconstruction with C4 symmetry was determined to an overall resolution of 28 Å using gold-standard refinement procedure (Extended Data Fig. 2).

Dose fractionated super-resolution image stacks of frozen, hydrated MBP–TRPA1 images collected using K2 Summit camera were first binned 2 × 2 resulting in a pixel size of 1.22 Å for motion correction and further image processing. Each image stack was subjected to whole-frame motion correction[53], and a sum of all subframes in each image stack was used for further processing. Particle picking was performed using a previously described procedure implemented in a Python script 'samautopick.py'. 2D class averages generated from manually picked particles were used as initial reference. All picked particles were then screened visually and bad particles identified in the visual screening were removed interactively. The selected particles were further screened by a reference-free 2D classification. The 28 Å resolution 3D reconstruction was low-pass filtered to a resolution of 60 Å, and used as the initial reference model for the 3D classification procedure

using RELION. Stable classes were then iteratively refined and reclassified to obtain the most homogeneous subset for the final 3D reconstruction. All refinements follow the gold-standard refinement procedure, in which the data set was divided into two half sets, and refined independently. Once the refinement is converged, the final data set was subjected to a movie processing and particle polishing procedure implemented in RELION[58]. A mask is generated to remove unstructured densities, such as those corresponding to the crescent and PMAL-C8, before calculating the final FSC curve. The final resolutions were estimated using the FSC = 0.143 criterion[59] on corrected FSC curves in which the influences of the mask were removed. Local resolution was estimated from unbinned and unsharpened raw density map using ResMap[60] within the RELION suite. The initial and final number of particles, as well as resolution for each data set is reported in Extended Data Table 1.

Model building and refinement. For *ab initio* model building in COOT[61], maps amplified with different negative temperature factors were used. Atomic model building was carried out mostly using the AITC-treated cryo-EM map and cross-validated with the single and double antagonists-treated cryo-EM maps. For the transmembrane domain, a homology model generated with HHPred[62] based on TRPV1 (PDB code 3J5P; ref. 21) was placed into the cryo-EM density map as a placeholder. For the ARD, we first generated sequence-based homology models of individual ankyrin repeats using secondary structure predictions[63], repeat prediction servers[64] and available literature[65]. These homology models were placed into the density map based on their position in the sequence, starting from the most C-terminal repeat. A total of five ankyrin repeats were properly placed into the well-resolved density of the ARD. Connections between these fitted ankyrin repeats were then built into the densities. For the remaining density, including the coiled-coil domain, linker between the last ankyrin repeat and the S1 helix, a polyalanine model was first built and amino acid assignment to the polyalanine model was then achieved based mainly on the clearly defined side chains densities of bulky residues. The entire model was then manually adjusted to fit the densities. Connectivity and side-chain rotamer positions were cross-validated using the two other maps. Clear densities were observed for residues 664–679, 748–763 and 786–802, but the resolutions of these domains are insufficient for atomic model building. The connection between the TRP-like domain and the coiled-coil domain is only partially resolved. At low σ isosurface, weak densities were seen to link the TRP-like domain to a short helical-like density in the inner leaflet of the membrane (Extended Data Fig. 7b) and then to the coiled-coil domain. We did not attempt to build a model for this linker region. The ~30 remaining most C-terminal residues were not resolved, probably forming part of the crescent density. The first 444 amino acids could not be modelled into the crescent density. Sequence-based prediction identified at least 11 additional ankyrin repeats. A speculative model was built by positioning a sequence based homology model of 11 ankyrin repeats symmetrically into the crescent density (Extended Data Fig. 8e).

After the initial model building and local energy minimization, the entire model was subject to reciprocal space refinement. Amplitudes of the final density map were corrected by a frequency-dependent scaling factor determined by comparing the experimental maps with a reference map calculated from the model[66], using a program 'ampcorr' written by X. Li. A soft-edged mask was generated based on the built atomic model to mask out densities of PMAL-C8, N-terminal crescent density, and other parts where model building was not attempted. The masked maps were put into an artificial unit cell with P1 symmetry and converted to MTZ format using CCP4 program sftools[67]. The resulting reflection files were used to perform maximum likelihood refinement using PHENIX[68] with secondary structure restraints, reference model restraints, and automatic optimization of experimental/stereochemistry weights. The reference model was generated from the built models using the geometry minimization function in PHENIX. The refined atomic model was further visualized in COOT. A few residues with side chain moving out of the density during the refinement were fixed manually, followed by further refinement following the same procedure. The final structure was validated using MolProbity[69].

For cross-validation to prevent overfitting, we followed procedures as previously described[70]. In brief, the coordinates of the refined atomic structures were displaced randomly by 0.2 Å using PHENIX (PDB tools) to remove potential model bias. The displaced model was then refined against one of the half maps in reciprocal space. FSC curves were calculated between the resulting model and half map 1 ('work', that is, used for refinement), the resulting model and half map 2 ('free', that is, not used for refinement), and the resulting model and the summed map (Extended Data Fig. 6). There is no significant separation between work and free FSC curves, suggesting that the model was not overfitted. The final model exhibits good geometry and all refinement parameters are listed in Extended Data Fig. 6.

Figures were prepared using UCSF Chimera[71], residue interactions were assigned with a 5 Å cut-off, and pore radii were calculated using the HOLE program[72].

50. Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14,** 673–681 (2006).
51. Chae, P. S. *et al.* Maltose-neopentyl glycol (MNG) amphiphiles for solubilization, stabilization and crystallization of membrane proteins. *Nature Methods* **7,** 1003–1008 (2010).
52. Ohi, M., Li, Y., Cheng, Y. & Walz, T. Negative staining and image classification - powerful tools in modern electron microscopy. *Biol. Proced. Online* **6,** 23–34 (2004).
53. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10,** 584–590 (2013).
54. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
55. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116,** 190–199 (1996).
56. Elmlund, H., Elmlund, D. & Bengio, S. PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. *Structure* **21,** 1299–1306 (2013).
57. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
58. Scheres, S. H. Beam-induced motion correction for sub-megadalton cryo-EM particles. *Elife* **3,** e03665 (2014).
59. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9,** 853–854 (2012).
60. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
61. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
62. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–W248 (2005).
63. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195–202 (1999).
64. Gruber, M., Soding, J. & Lupas, A. N. REPPER–repeats and their periodicities in fibrous proteins. *Nucleic Acids Res.* **33,** W239–W243 (2005).
65. Gaudet, R. A primer on ankyrin repeat function in TRP channels and beyond. *Mol. Biosyst.* **4,** 372–379 (2008).
66. Penczek, P., Ban, N., Grassucci, R. A., Agrawal, R. K. & Frank, J. *Haloarcula marismortui* 50S subunit-complementarity of electron microscopy and X-Ray crystallographic information. *J. Struct. Biol.* **128,** 44–50 (1999).
67. Winn, M. D. *et al.* Overview of the *CCP*4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
68. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
69. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
70. Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343,** 1485–1489 (2014).
71. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
72. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14,** 354–360, 376 (1996).

LETTER

Self–similar fragmentation regulated by magnetic fields in a region forming massive stars

Hua-bai Li[1], Ka Ho Yuen[1], Frank Otto[1], Po Kin Leung[1], T. K. Sridharan[2], Qizhou Zhang[2], Hauyu Liu[3], Ya-Wen Tang[3] & Keping Qiu[4]

Most molecular clouds are filamentary or elongated[1–3]. For those forming low-mass stars (<8 solar masses), the competition between self-gravity and turbulent pressure along the dynamically dominant intercloud magnetic field (10 to 100 parsecs) shapes the clouds to be elongated either perpendicularly[4] or parallel[5] to the fields. A recent study[6] also suggested that on the scales of 0.1 to 0.01 parsecs, such fields are dynamically important within cloud cores forming massive stars (>8 solar masses). But whether the core field morphologies are inherited from the intercloud medium or governed by cloud turbulence is unknown, as is the effect of magnetic fields on cloud fragmentation at scales of 10 to 0.1 parsecs[7–9]. Here we report magnetic-field maps inferred from polarimetric observations of NGC 6334, a region forming massive stars, on the 100 to 0.01 parsec scale. NGC 6334 hosts young star-forming sites[10–12] where fields are not severely affected by stellar feedback, and their directions do not change much over the entire scale range. This means that the fields are dynamically important. The ordered fields lead to a self-similar gas fragmentation: at all scales, there exist elongated gas structures nearly perpendicular to the fields. Many gas elongations have density peaks near the ends, which symmetrically pinch the fields. The field strength is proportional to the 0.4th power of the density, which is an indication of anisotropic gas contractions along the field.

We conclude that magnetic fields have a crucial role in the fragmentation of NGC 6334.

At a distance of about 1.7 kiloparsecs from Earth, NGC 6334 is one of the nearest regions forming massive stars. Other sites forming massive stars are usually too far away to use starlight polarization (owing to extinction from dust grains aligned by magnetic fields, *B*-fields) effectively to probe the *B*-field orientations (the starlight polarization is parallel to the orientation of the *B*-field) in the surrounding intercloud medium. *B*-field directions can be derived from the polarization of local background stars by subtracting the polarization of local foreground stars[13]. To form a typical giant molecular cloud, gas needs to be accumulated from an intercloud medium of a few hundred parsecs[14]. Using the optical polarimetry archive from ref. 15, the ambient *B*-field direction of NGC 6334 is seen to be perpendicular to its elongation (Fig. 1).

If the *B*-field is dynamically important compared to turbulence during the gas accumulation process, the ambient *B*-field direction should be preserved inside the cloud[8]. Regions forming massive stars have stronger thermal dust emissions than do those forming low-mass stars, which allows probing of the *B*-fields within dense clouds using polarization of submillimetre thermal dust emissions (the two directions are perpendicular). This has been performed extensively[13,16,17], but never with
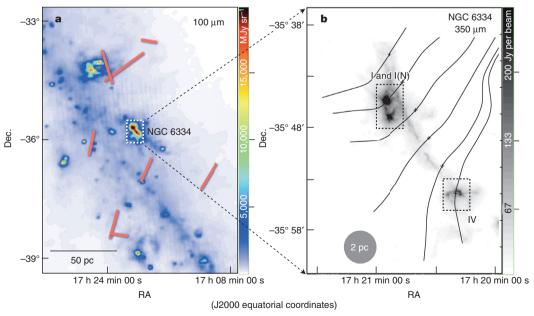
Figure 1 | The *B*-field directions of NGC 6334 and the local intercloud medium. a, *B*-field directions (red) inferred from optical polarimetry[15] overlapped with the 100-μm map from IRAS. **b**, Zooming in on **a**, SPARO[13] showed the field lines inferred from the 450-μm polarimetry with a resolution of 2 pc overlapped with a 350-μm map[30]. The filamentary cloud 'pinches' the field lines and the intensity peaks at the two ends of the filament (dashed rectangles), where the *B*-field morphologies with higher resolutions are shown in Fig. 2. The top two field lines are used to estimate the field curvature. N, north.

[1]Department of Physics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong, China. [2]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. [3]Academia Sinica Institute of Astronomy and Astrophysics, 11F Astronomy-Mathematics Building, AS/NTU (National Taiwan University) No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan, China. [4]School of Astronomy and Space Science, Nanjing University, 22 Hankou Road, Nanjing, Jiangsu 210093, China.
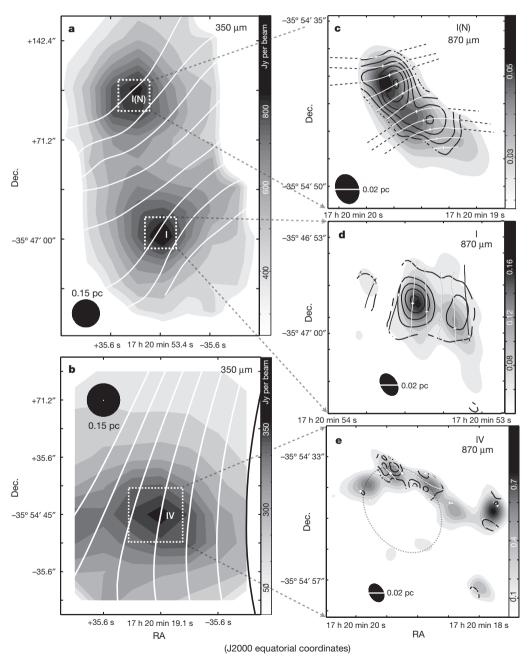
such a wide range in scales as presented here. We use data acquired by the polarimeters SPARO (Submillimeter Polarimeter for Antarctic Remote Observing) (10-pc scale)[13], Hertz (1-pc scale)[18], and the Submillimetre Array (SMA, 0.1-pc scale)[6]. Following ref. 19, interpolation of independent 3σ polarimetry detections is used to plot the B-field lines (Figs 1 and 2).

It is obvious that the field lines at the 10-pc scale are 'pinched' near the ends of the dust filament, where the massive-star-forming clumps, I/I(N) and IV, are also located (Fig. 1; N denotes north). Hertz resolved the density peaks, showing that I and I(N) are again situated near field-line pinches at the parsec scale (Fig. 2a).

SMA further zoomed in onto the density peaks of I, I(N), and IV. I(N) is the youngest of the three cores, with weak outflows[6,10] and a low

temperature ($\sim$30 K)[11]; the field lines are again symmetrically pinched (Fig. 2c). The more developed core I is hotter ($\sim$100 K)[12], and has high-velocity outflows in the northeast–southwest direction[10], which might have altered the field direction from the larger scale. The curved filament of IV is part of a compression shell due to the H II bubble[20], which compressed the B-fields at the same time; hence the field and filament are largely aligned (Fig. 2e).

The average orientations of the filamentary cloud, elongated clumps/cores and the B-fields (defined by "equal weight Stokes mean"[13]) are summarized in Fig. 3. The orientation of a cloud is defined by the long-axis direction of the autocorrelation function of the intensity map[3]. There are several intriguing facts revealed by Fig. 3. First, assuming turbulence is the only force that disturbs B-field orientations and has the

Figure 2 | The B-field within the clumps/cores. a, b, Clumps observed by Hertz/Caltech Submillimeter Observatory (CSO)[18] with a 0.15-pc resolution. The intensity again peaks near the field line pinches in **a**. The four field lines passing through the dashed rectangles in **a** are used to estimate the field curvature. **c–e,** Cores observed by SMA[6] with a 0.02-pc resolution. Some field lines in **c** are extended (as dashed lines to help visualize the pinches) and are used to estimate the curvature. In **e**, the oval indicates a shell H II region[20]. The contours show the relative intensity of the polarized flux, which tends to increase with the total intensity.
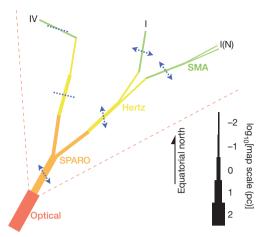
Figure 3 | Self-similar fragmentation and field configurations at 100–0.01 pc. Each solid line shows the mean field direction within a map, whose scale is indicated by the linewidth (key at bottom right). The blue dashed lines show the cloud long-axis directions. At the ends of a dashed line, arrowheads are added if density peaks at the ends of the cloud, where the field directions are indicated by the branched lines. The red dashed lines deviate from the mean intercloud-medium field (optical) by 30°. Except for core IV, B-field directions vary within the range defined by the red dashed lines.

same energy density as the B-fields, the dispersion of B-field directions should be 30° based on the Chandrasekhar–Fermi relation[3,21,22]. Apart from region IV, all the field orientations in Fig. 3 are within this 30° range. In reality, field dispersions are not only due to turbulence, but also gravity[16,17], stellar feedback (for example, region IV) and projection, so the turbulent energy of NGC 6334 should be sub-Alfvenic (see Methods section 'B-field direction alignment between scales').

Second, at all the scales, we observed hourglass-shaped or ordered B-fields to be close to perpendicular to cloud elongations unless severely affected by stellar feedback (core IV). This is a signature of the Lorentz force supporting the cloud against gravitational contraction in the direction perpendicular to the field lines[16,17]. This anisotropic contraction will result in flattened structures, which will appear elongated and tend to be perpendicular to the B-field projection[3]. Third, a thin sheet should fragment at the rim instead of at the centre of the mass because of the difference in gravitational contraction velocity over the sheet[23]. The sky projection should appear as off-centred density peaks near the field-line pinches, which is also observed at multiple scales.

How B-field strengths vary with gas density also reveals the role of B-fields in the contraction of cloud fragments. If B-fields are dynamically unimportant, the contraction should be isotropic, which results in $B \propto n^{2/3}$, where n is density. This is because contraction along the B-field direction can enhance n but not B. The exponent should be less than 2/3 if the B-fields are strong enough to channel the contraction to some extent[24,25]. Since we cover multiple scales and thus multiple densities, we can study this dependence and its implications.

We can estimate the B-field strength based on the balance between the forces from gravity (F_G), magnetic pressure (F_P) and magnetic tension (F_T) (Fig. 4; see Methods section 'Estimate of field strength'). The F_G between the density peaks is:
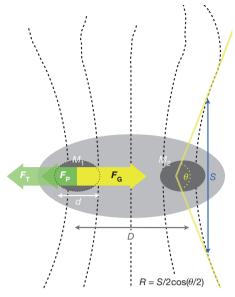
$$F_G = 2.8 \times 10^{28} \left(\frac{M_1}{100 M_\odot} \right) \left(\frac{M_2}{100 M_\odot} \right) \left(\frac{0.1 \text{ pc}}{D} \right)^2 \text{ dyn}$$

where M_1 and M_2 are the masses of the two dense clumps; $M_\odot$ is the solar mass; and D is the distance between the peaks.

Presenting B-field orientations as field lines[19] allows us to estimate field-line curvatures and thus F_T:

$$F_T = \frac{V}{4\pi} B \cdot \nabla B \approx \frac{V}{4\pi} \frac{B^2}{R} = 1.5 \times 10^{30} \left(\frac{B}{1 \text{ mG}} \right)^2 \left(\frac{0.5 \text{ pc}}{R} \right) \left(\frac{V}{1 \text{ pc}^3} \right) \text{ dyn}$$

where R is the radius of the field-line curvature and V is the clump

Figure 4 | Parameters used to estimate B-field strength. M_1 and M_2 are the clump masses. d is the clump size (for estimating V). D is the distance between clumps. R is the field-line (dashed line) curvature radius. From the tip of a pinch, moving in both directions along the field line, one can draw tangents (yellow lines), which form an angle θ. We stop moving the tangent points when θ stops decreasing. R can then be derived from the equation shown, where S is the separation between the two tangent points. F_T, F_P, and F_G represent the forces due to magnetic tension, magnetic pressure, and gravity, respectively.

volume. Field lines near the density peaks and with more prominent curvatures are selected (noted in Figs 1 and 2) to estimate R. This will give a lower limit for R (and thus B).

If there exists a gradient of the B-field strength, F_P should also be considered:

$$F_P = -\nabla \left(\frac{B^2}{8\pi} \right) V \approx \frac{V}{8\pi} \frac{B^2 - B_0^2}{r/2}$$

$$= 1.5 \times 10^{30} \left[\left(\frac{B}{1 \text{ mG}} \right)^2 - \left(\frac{B_0}{1 \text{ mG}} \right)^2 \right] \left(\frac{0.5 \text{ pc}}{r} \right) \left(\frac{V}{1 \text{ pc}^3} \right) \text{ dyn}$$

where B_0 is the field strength outside the hourglass and r is twice the 'waist' (approximated by D). At the 10-pc scale, the Galactic field strength, 10 μG (ref. 25), is used for B_0. Estimates of B on the 10-pc and 1-pc scales are used as B_0 for the 1-pc and 0.1-pc scales respectively. M, D and V can be derived from the literature[26,27] (Extended Data Table 1) and R is measured from our maps (Figs 1, 2 and 4).

Setting $F_G = F_P + F_T$ (see Method section 'Estimate of B-field strength') results in B of approximately 0.2 mG, 1.2 mG and 12 mG on the 10-pc, 1-pc and 0.1-pc scales respectively (Extended Data Table 1). Note that the line of sight (LOS) and the Sagittarius spiral arm (and thus the Galactic B-field) are almost perpendicular at the position of NGC 6334. Hence the projection effect on the field curvatures and thus strengths should not be severe. The LOS component of B at the parsec scale measured by the Zeeman effect is roughly 0.2 mG (ref. 28), which implies that the angle between the B-field and LOS is 80°.

Finally, approximating n by $(M_1 + M_2)/D^3$ yields $B \propto n^{0.41 \pm 0.04}$ (where $0.04 \sim 2\sigma$; Extended Data Fig. 1), with an exponent $\sim 13\sigma$ below 2/3. This is the first B–n relation derived from one single cloud covering 10–0.1 pc. Previously, the exponents are mainly based on Zeeman measurements where different values of n are obtained from different types of clouds that do not have any connection with each other[24,25]. We can further show that the ratios of mass to magnetic flux of the cloud/clump/cores are on average 1.6 ± 0.5 relative to the critical value, the ratio at which self-gravity balances magnetic forces (see

Methods section 'Mass-to-flux ratio'). This agrees with the value required to form massive stars in recent numerical studies[9].

The magnetic topology problem[29], that is, how the field topology evolves as molecular clouds form out of the interstellar medium and as cores contract to form stars, has puzzled astronomers for decades, largely owing to the difficulties of observation. After a decade of data collection, we finally shed some light on this problem. The Atacama Large Millimeter/submillimetre Array will have adequate sensitivity/resolution to survey *B*-fields in young massive-star-forming sites beyond NGC 6334.

1. André, Ph. *et al.* From filamentary clouds to prestellar cores to the stellar IMF: initial highlights from the Herschel Gould Belt Survey. *Astron. Astrophys.* **518,** L102–L108 (2010).
2. Zhang, Q., Wang, Y., Pillai, T. & Rathborne, J. Fragmentation at the earliest phase of massive star formation. *Astrophys. J.* **696,** 268–273 (2009).
3. Li, H.-b., Fang, M., Henning, T. & Kainulainen, J. The link between magnetic fields and filamentary clouds: bimodal cloud orientations in the Gould Belt. *Mon. Not. R. Astron. Soc.* **436,** 3707–3719 (2013).
4. Nakamura, F. & Li, Z.-Y. Magnetically regulated star formation in three dimensions: the case of the Taurus Molecular Cloud Complex. *Astron. J.* **687,** 354–375 (2008).
5. Stone, J., Ostriker, E. & Gammie, C. Dissipation in compressible magnetohydrodynamic turbulence. *Astrophys. J.* **508,** L99–L102 (1998).
6. Zhang, Q. *et al.* Magnetic fields and massive star formation. *Astrophys. J.* **792,** 116–127 (2014).
7. Collins, D. *et al.* The two states of star-forming clouds. *Astrophys. J.* **750,** 13–30 (2012).
8. Li, H.-b., Dowell, C., Goodman, A., Hildebrand, R. & Novak, G. Anchoring magnetic field in turbulent molecular clouds. *Astrophys. J.* **704,** 891–897 (2009).
9. Myers, A., McKee, C., Cunningham, A., Klein, R. & Krumholz, M. The fragmentation of magnetized, massive star-forming cores with radiative feedback. *Astrophys. J.* **766,** 97–114 (2013).
10. McCutcheon, W. *et al.* Star formation in NGC 6334 I and I(N). *Mon. Not. R. Astron. Soc.* **316,** 152–164 (2000).
11. Beuther, H. *et al.* Hot ammonia in NGC 6334I & I(N). *Astron. Astrophys.* **466,** 989–998 (2007).
12. Sandell, G. (Sub)mm continuum mapping of NGC 6334 I & I(N). A cobweb of filaments and protostars. *Astron. Astrophys.* **358,** 242–256 (2000).
13. Li, H.-b. *et al.* Results of SPARO 2003: mapping magnetic fields in giant molecular clouds. *Astrophys. J.* **648,** 340–354 (2006).
14. Williams, J., Blitz, L. & McKee, C. in *Protostars and Planets IV* (eds Mannings, A., Boss, P. & Russell, S.) 97–111 (Univ. Arizona Press, 2000).
15. Heiles, C. 9286 stars: an agglomeration of stellar polarization catalogs. *Astron. J.* **119,** 923–927 (2000).
16. Schleuning, D. Far-infrared and submillimeter polarization of OMC-1: evidence for magnetically regulated star formation. *Astrophys. J.* **493,** 811–825 (1998).
17. Girart, J., Rao, R. & Marrone, D. Magnetic fields in the formation of Sun-like stars. *Science* **313,** 812–814 (2006).
18. Dotson, J. L. *et al.* 350 μm Polarimetry from the Caltech Submillimeter Observatory. *Astrophys. J.* **186** (Suppl.), 406–426 (2010).
19. Li, H.-b., Houde, M., Lai, S.-p. & Sridharan, T. K. Tracing turbulent ambipolar diffusion in molecular clouds. *Astrophys. J.* **718,** 905–912 (2010).
20. Rodríguez, L. *et al.* Radio detection of the exciting sources of shell H II regions in NGC 6334. *Rev. Mex. Astron. Astrofís.* **15,** 194–196 (2003).
21. Chandrasekhar, S. & Fermi, E. Magnetic fields in spiral arms. *Astrophys. J.* **118,** 113–115 (1953).
22. Falceta-Gonçalves, D., Lazarian, A. & Kowal, G. Studies of regular and random magnetic fields in the ISM: statistics of polarization vectors and the Chandrasekhar-Fermi technique. *Astrophys. J.* **679,** 537–551 (2008).
23. Burkert, A. & Hartmann, L. Collapse and fragmentation in finite sheets. *Astrophys. J.* **616,** 288–300 (2004).
24. Li, H.-b. *et al.* in *Protostars and Planets VI* (eds Beuther, H., Klessen, R., Dullemond, C. & Henning, Th.) 101–123 (Univ. Arizona Press, 2014).
25. Crutcher, R., Wandelt, B., Heiles, C., Falgarone, E. & Troland, T. Magnetic fields in interstellar clouds from Zeeman observations: inference of total field strengths by Bayesian analysis. *Astrophys. J.* **725,** 466–479 (2010).
26. Matthews, H., McCutcheon, W., Kirk, H., White, G. & Cohen, M. The distribution and properties of cold dust in NGC 6334. *Astron. J.* **136,** 2083–2101 (2008).
27. Hunter, T. *et al.* Millimeter multiplicity in NGC 6334 I and I(N). *Astrophys. J.* **649,** 888–893 (2006).
28. Sarma, A., Troland, T., Roberts, D. & Crutcher, R. VLA OH and H I Zeeman observations of the NGC 6334 complex. *Astrophys. J.* **533,** 271–280 (2001).
29. McKee, C., Zweibel, E. & Goodman, A. and Heiles, C. Magnetic fields in star-forming regions—theory. In *Protostars and Planets III* (eds Levy, E. & Lunine, J.) 327–366 (1993).
30. Vaillancourt, J. Observations of polarized dust emission at far-infrared through millimeter wavelengths. *Astron. Soc. Pacif. Conf. Ser.* **449,** 169–181 (2011).

Author Contributions H.-b.L. designed and executed the experiment. K.H.Y. measured the field curvatures. F.O. performed the numerical simulations. The Chinese University of Hong Kong team was responsible for the manuscript. The CfA-ASIAA-Nanjing team helped with the SMA data acquisition and reduction.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.-b.L. (hbli@phy.cuhk.edu.hk).

METHODS

B-field direction alignment between scales. Note that the *B*-field dispersion test on the relative strength between turbulence and the *B*-field (Fig. 3) might not work in the close vicinity of stars, where gas rotation and/or stellar feedback govern the field orientations. This may also be why protostellar disks are found to align with fields at the 10^{-2}-pc scale (poloidal)[31] but not at the 10^{-3}-pc scale[32]. The transformation from poloidal (cores) to toroidal (discs) fields[33] can orient *B*-fields in any direction at the 10^{-3}-pc scale.

Estimate of *B*-field strength. *The assumption* of $F_G = F_P + F_T$. We assumed that self-gravity and magnetic fields are close to virial equilibrium, the so-called critical condition. In practice, we use force equilibrium, $F_G = F_P + F_T$, instead of virial equilibrium to estimate field strengths; the latter involves volume integration over the cloud. In Methods section 'Mass-to-flux ratio', we show that our result is roughly consistent with virial equilibrium. Here we explain why the cloud should be close to the critical condition.

First, given the hourglass-shaped field morphologies, self-gravity is able to compress the fields. So the cloud cannot be considerably subcritical ($F_G \ll F_P + F_T$). Second, given the elongation of the cores and clumps, the cloud cannot be considerably supercritical ($F_G \gg F_P + F_T$), either. A spherical contraction happens only when the gas is significantly supercritical and it should not be difficult to appreciate that the stronger the field is, the more elongated the clumps/cores should be. To study the relation between cloud elongation and criticality, we used the ZEUS-MP code[34] to simulate cloud contraction. We designed the simulations to be as simple as possible to focus on the interaction between self-gravity and *B*-fields. The initial condition is a uniform spherical cloud embedded in a uniform *B*-field, with negligible gas pressure, no turbulence and no ambipolar diffusion. The only variable is the *B*-field strength, such that the mass-to-flux ratio (MFR) is 1, 2, 4 and 8 times the critical value[35]. The short-to-long axis ratios obtained after ten million years of contraction are shown in Extended Data Fig. 2. We note that it does not take much super-criticality for a nearly spherical contraction; MFR > 4 is enough. Extended Data Fig. 2 also displays the measured axis ratios of the clumps or cores in NGC 6334; indeed they are found to be close to the critical condition.

Third, recent studies[3,24] compared the empirical threshold column density of cloud contraction[36,37] with the magnetic critical column density[24,25] of the typical Galactic field ($\sim 10\,\mu G$) and found a very good agreement. For densities lower than this threshold, the field strength is independent of density[24,25], that is, gas is accumulated along field lines. This is consistent with the scenario of fragmentation channelled by *B*-fields: subcritical gas is accumulated along the field lines till the cloud becomes critical and able to compress the field lines (pinches)[3,24]. The cloud in Fig. 1b indeed looks much more elongated compared to the critical condition in Extended Data Fig. 2. Moreover, owing to flux freezing, field-line compression will not change the magnetic criticality[38], and thus magnetic virial equilibrium should be a good approximation for all scales. In this picture, if Zeeman measurements are used to estimate field strengths, we should expect clouds to range from just critical to highly supercritical due to the projection effect (because only the LOS components are detected by Zeeman measurements). This range is indeed observed by ref. 25. We emphasize that this interpretation of the Zeeman measurements is debated; other authors[25] interpret the observed range as an indication that some cores can be highly supercritical. Highly supercritical cores, however, are not supported by the surveys mentioned above showing that the contraction threshold agrees well with the magnetic critical column density[3,24]. Moreover, it is now understood that Zeeman measurements can potentially underestimate mean field strengths[24] owing to B_{LOS} reversals within a telescope beam[39,40,41,42]. In any case, filaments like NGC 6334 should not belong to the highly supercritical category even if there is one, because of its low short-to-long axis ratio.

Comparison with the Chandrasekhar–Fermi method. Attributing all the field structures to turbulence, Chandrasekhar and Fermi[3,21,22] proposed that field strength should be estimated as follows:

$$B = \frac{1}{2}\sqrt{4\pi\rho}\,\frac{\delta v}{\delta \alpha}\ \text{gauss}$$

where ρ is gas density (in grams per centimetre cubed), δv is the LOS velocity dispersion (in centimetres per second) and $\delta \alpha$ is the *B*-field direction angle dispersion (in radians); the factor 1/2 is a correction suggested by numerical simulations[43]. In Fig. 2a, clump I/I(N), as an example, $\delta \alpha$ is measured as $17.5°$ (ref. 44). The full-width half-maximum line width of CO(2–1) emission is detected as $13.7\ \text{km s}^{-1}$ for core I and $12.1\ \text{km s}^{-1}$ for core I(N) at the 26″ scale (the beam size)[45]. Using the separation

between cores I and I(N), 106.8″, as an estimate of the clump size, the linewidth at the clump scale can be estimated by: $\left(\frac{13.7 + 12.1}{2}\right)\left(\frac{106.8}{26}\right)^{0.5} = 26.1\ \text{km s}^{-1}$, assuming a 0.5 exponent for the turbulent velocity spectrum[46]. Converting the linewidth to the velocity dispersion gives $\delta v = 11.1\ \text{km s}^{-1}$. Assuming $n(H_2) = 10^4\ \text{cm}^{-3}$ and a mean molecular mass of 2.8, the above equation gives $B = 1.4\ \text{mG}$, which is comparable to our estimate of $1.2 \pm 0.7\ \text{mG}$ (Extended Data Table 1). Note that apparently gravity also plays a part in determining the field structure of Fig. 2a (the hourglass shape), so the estimate from the Chandrasekhar–Fermi method should be a lower limit.

Mass-to-flux ratio. From Extended Data Table 1, we can roughly estimate the MFR, which is familiar to astronomers when comparing gravitational and magnetic forces. The critical MFR (that is, when self-gravity and magnetic fields reach virial equilibrium) is sensitive to cloud geometries; for example, $1/2\pi\sqrt{G}$ for a disk[35] and $2/3\pi\sqrt{G}$ for a spherical cloud[47], where G is the gravitational constant. Assuming D from Extended Data Table 1 to be the cross-sectional diameters, the MFRs normalized to the critical value are approximately 1.1 ± 0.24, 2.4 ± 0.74, 2.2 ± 0.54 and 1.7 ± 0.32 for the cloud, clump I/I(N), core I(N) and core I, respectively, based on the equation from ref. 35; correspondingly, the values are 0.83 ± 0.18, 1.8 ± 0.55, 1.7 ± 0.41 and 1.3 ± 0.25 based on ref. 47. The shapes of our objects fall between a disk and a sphere, and the average of the above values is 1.6 ± 0.5. Although the approximations of the cloud shapes and cross-sections are rough, the cloud is unlikely to be highly supercritical.

We can check the consistency between the observed MFR and *B*–*n* relation using the same sets of simulations discussed in Methods section 'Estimate of field strength'. The *B*–*n* relations for the first ten million years are shown in Extended Data Fig. 2. The 0.4-power indeed occurs when the MFR is 1 to 2 times of the critical value, which is consistent with our observation.

31. Chapman, N. *et al.* Alignment between flattened protostellar infall envelopes and ambient magnetic fields. *Astrophys. J.* **770,** 151–165 (2013).
32. Hull, C. *et al.* Misalignment of magnetic fields and outflows in protostellar cores. *Astrophys. J.* **768,** 159–163 (2013).
33. Tomisaka, K. Origin of molecular outflow determined from thermal dust polarization. *Publ. Astron. Soc. Jpn.* **63,** 147–158 (2011).
34. Hayes, J. *et al.* Simulating radiating and magnetized flows in multiple dimensions with ZEUS-MP. *Astrophys. J. Suppl. Ser.* **165,** 188–228 (2006).
35. Nakano, T. & Nakamura, T. Gravitational instability of magnetized gaseous disks 6. *Publ. Astron. Soc. Jpn* **30,** 671–680 (1978).
36. Kainulainen, J., Beuther, H., Henning, T. & Plume, R. Probing the evolution of molecular cloud structure. From quiescence to birth. *Astron. Astrophys.* **508,** L35–L38 (2009).
37. Froebrich, D. & Rowles, J. The structure of molecular clouds—II. Column density and mass distributions. *Mon. Not. R. Astron. Soc.* **406,** 1350–1357 (2010).
38. Elmegreen, B. On the rapid collapse and evolution of molecular clouds. *Astrophys. J.* **668,** 1064–1082 (2007).
39. Bertram, E., Federrath, C., Banerjee, R. & Klessen, R. Statistical analysis of the mass-to-flux ratio in turbulent cores: effects of magnetic field reversals and dynamo amplification. *Mon. Not. R. Astron. Soc.* **420,** 3163–3173 (2012).
40. Lunttila, T., Padoan, P., Juvela, M. & Nordlund, A. The Super-Alfvénic model of molecular clouds: predictions for Zeeman splitting measurements. *Astrophys. J.* **686,** L91–L94 (2008).
41. Mouschovias, T. & Tassis, K. Testing molecular-cloud fragmentation theories: self-consistent analysis of OH Zeeman observations. *Mon. Not. R. Astron. Soc.* **400,** 15–19 (2009).
42. Bourke, T., Myers, P., Robinson, G. & Hyland, A. New OH Zeeman measurements of magnetic field strengths in molecular clouds. *Astrophys. J.* **554,** 916–932 (2001).
43. Ostriker, E., Stone, J. & Gammie, C. Density, velocity, and magnetic field structure in turbulent molecular cloud models. *Astrophys. J.* **546,** 980–1005 (2001).
44. Novak, G., Dotson, J. L. & Li, H. Dispersion of observed position angles of submillimeter polarization in molecular clouds. *Astrophys. J.* **695,** 1362–1369 (2009).
45. Kraemer, K. & Jackson, J. Molecular gas in the NGC 6334 star formation region. *Astrophys. J.* **124** (Suppl.), 439–463 (1999).
46. Heyer, M. & Brunt, C. The universality of turbulence in galactic molecular clouds. *Astrophys. J.* **615,** 45–48 (2004).
47. Strittmatter, P. Gravitational collapse in the presence of a magnetic field. *Mon. Not. R. Astron. Soc.* **132,** 359–378 (1966).
48. Chibueze, J. *et al.* Astrometry and spatio-kinematics of H$_2$O masers in the massive star-forming region NGC 6334I(North) with VERA. *Astrophys. J.* **784,** 114–121 (2014).
49. Reid, M. *et al.* Trigonometric parallaxes of high mass star forming regions: the structure and kinematics of the Milky Way. *Astrophys. J.* **783,** 130–143 (2014).

LETTER

Phonon counting and intensity interferometry of a nanomechanical resonator

Justin D. Cohen[1]*, Seán M. Meenehan[1]*, Gregory S. MaccCabe[1], Simon Gröblacher[1,2]†, Amir H. Safavi-Naeini[1,3], Francesco Marsili[4], Matthew D. Shaw[4] & Oskar Painter[1]

In optics, the ability to measure individual quanta of light (photons) enables a great many applications, ranging from dynamic imaging within living organisms[1] to secure quantum communication[2]. Pioneering photon counting experiments, such as the intensity interferometry performed by Hanbury Brown and Twiss[3] to measure the angular width of visible stars, have played a critical role in our understanding of the full quantum nature of light[4]. As with matter at the atomic scale, the laws of quantum mechanics also govern the properties of macroscopic mechanical objects, providing fundamental quantum limits to the sensitivity of mechanical sensors and transducers. Current research in cavity optomechanics seeks to use light to explore the quantum properties of mechanical systems ranging in size from kilogram-mass mirrors to nanoscale membranes[5], as well as to develop technologies for precision sensing[6] and quantum information processing[7,8]. Here we use an optical probe and single-photon detection to study the acoustic emission and absorption processes in a silicon nanomechanical resonator, and perform a measurement similar to that used by Hanbury Brown and Twiss to measure correlations in the emitted phonons as the resonator undergoes a parametric instability formally equivalent to that of a laser[9]. Owing to the cavity-enhanced coupling of light with mechanical motion, this effective phonon counting technique has a noise equivalent phonon sensitivity of 0.89 ± 0.05. With straightforward improvements to this method, a variety of quantum state engineering tasks using mesoscopic mechanical resonators would be enabled[10], including the generation and heralding of single-phonon Fock states[11] and the quantum entanglement of remote mechanical elements[12,13].

Measurement of the properties of mechanical systems in the quantum regime typically involves heterodyne detection of a coupled optical or electrical field, yielding a continuous signal proportional to the displacement amplitude[14]. An alternative method, particularly suited to optical read-out, is to utilize photon counting as a means to probe the quantum dynamics of the coupled optomechanical system[15,16]. Photon counting can be readily adapted to study intensity correlations in an optical field, and has been used not only in the astronomical studies of thermal light using the technique of Hanbury Brown and Twiss (HBT), but also in early studies of the photon statistics of laser light and single-atom fluorescence[4,17]. In the field of photon-correlation spectroscopy, such intensity interferometry techniques have found widespread application in the measurement of particle and molecular motion in materials[18]. More recently, photon counting of Raman scattering events in diamond has heralded and verified the quantum entanglement of a terahertz phonon shared between two separate bulk diamond crystals[13]. In the case of engineered cavity optomechanical systems, much longer phonon coherence times are attainable, albeit at lower mechanical frequencies (megahertz to gigahertz), which limit the temperature of operation

and the optical power handling capability of such structures. Quantum optical schemes for manipulation of the quantum state of motion in cavity optomechanical systems thus rely on a large per-phonon scattering rate and efficient detection of scattering events. Here we embed a high-Q, gigahertz-frequency mechanical resonator inside an optical nanocavity, greatly enhancing the phonon–photon coupling rate and channelling optical scattering into a preferred optical mode for collection. Single-photon detection of this scattered light then allows for a precise counting of single-phonon emission or absorption events, effectively phonon counting (although this terminology should not be confused with Fock state detection or quantum non-demolition measurement of phonon number). The highly engineered and optimized nature of this optomechanical resonator furthermore yields a sub-phonon-level counting sensitivity of the intracavity mechanical resonator occupancy.

A conceptual schematic of the phonon counting experiment is shown in Fig. 1a. The device consists of a patterned silicon nanobeam which forms an optomechanical crystal (OMC)[19,20] able to co-localize acoustic (mechanical) and optical resonances at frequencies ω_m and ω_c, respectively. Finite-element-method simulations of the acoustic and optical resonances are shown at the top of Fig. 1a. The Hamiltonian describing the interaction between the acoustic and optical modes is given by $\hat{H}_{int} = \hbar g_0 \hat{a}^\dagger \hat{a}(\hat{b} + \hat{b}^\dagger)$, where $\hat{a}(\hat{b})$ is the annihilation operator for the optical (acoustic) mode, and g_0 is the optomechanical coupling rate, physically representing the optical frequency shift due to the zero-point motion of the acoustic resonator. This interaction modulates a laser probe with frequency ω_l to produce sidebands at frequencies $\omega_l \pm \omega_m$, analogous to the anti-Stokes and Stokes sidebands in Raman scattering and corresponding to phonon absorption or emission, respectively. For a system in the resolved sideband limit, where $\omega_m \gg \kappa$ (κ is the linewidth of the optical resonance), the density of states of the optical cavity can be used to resonantly enhance either scattering process for an appropriately detuned pump. In particular, applying a large coherent pump red (blue) detuned from the optical cavity resonance by $\Delta = \omega_c - \omega_l = \omega_m$ ($\Delta = -\omega_m$) results in an effective interaction Hamiltonian of the form $\hat{H}_{int} = \hbar G(\hat{a}^\dagger \hat{b} + \hat{a}\hat{b}^\dagger)$ ($\hat{H}_{int} = \hbar G(\hat{a}\hat{b} + \hat{a}^\dagger \hat{b}^\dagger)$), where $G = g_0\sqrt{n_c}$ is the parametrically enhanced optomechanical coupling rate (n_c is the intracavity photon number at frequency ω_l due to the pump laser). In this case, the output field annihilation operator $\hat{a}_{out}$ can be shown to consist of a coherent component at frequency ω_l as well as a component at frequency ω_c which is proportional to $\hat{b}(\hat{b}^\dagger)$(ref. 21). Sending the cavity output through a series of narrowband optical filters centred on the cavity resonance, as shown in Fig. 1a, suppresses the pump so that photon counting events will correspond directly to counting phonon absorption (emission) events[22]. Subsequently directing the filter output to an HBT set-up in order to measure the second-order photon correlation function $g^{(2)}(\tau)$ (ref. 3) will then result in a direct measurement

[1]Institute for Quantum Information and Matter and Thomas J. Watson Senior Laboratory of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA. [2]Vienna Center for Quantum Science and Technology (VCQ), Faculty of Physics, University of Vienna, A-1090 Wien, Austria. [3]Edward L. Ginzton Laboratory, Stanford University, Stanford, California 94305, USA. [4]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. †Present address: Kavli Institute of Nanoscience, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands.
*These authors contributed equally to this work.
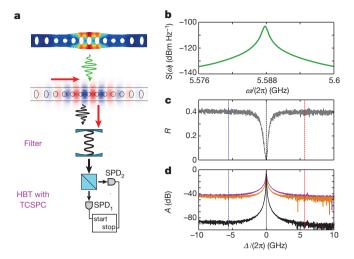
Figure 1 | Phonon counting and device characterization. a, Schematic of the phonon counting measurement. The finite element method simulations depict the displacement field of the acoustic resonance (top, blue bar) and the electric field of the optical resonance (next down, bar with no colour filling) of the nanobeam structure. Pump light at optical detuning $\Delta = \pm\omega_m$ is indicated by the red arrows, while the optomechanically scattered sideband light is represented by black arrows. The optical cavity output is filtered to reject the pump, and then detected in a Hanbury Brown and Twiss (HBT) set-up using two superconducting single photon detectors ($SPD_{1,2}$). The detector outputs are used as start/stop pulses in a time-correlated single photon counting module (TCSPC), yielding the second-order phonon correlation function. **b**, Measured power spectral density $S(\omega)$ of the acoustic resonance. **c**, Normalized optical cavity reflection spectrum, R. Pump detunings of $\Delta = \pm\omega_m/(2\pi) = \pm5.6$ GHz are indicated by the red and blue dashed lines, respectively. **d**, Transmission spectrum of the first (purple) and second (orange) optical filter, with total filter transmission plotted in black. A is the pump attenuation factor.

of the normally (anti-normally) ordered second-order phonon correlation function. Although this work deals with measurements of a particular cavity optomechanical system, the nanobeam OMC, numerous other geometries possess the requisite optomechanical coupling strength and high mechanical frequency necessary to implement this phonon counting scheme[5].

As described in ref. 20, the nanobeam is patterned in such a way as to support a 'breathing' acoustic resonance at $\omega_m/2\pi = 5.6$ GHz as well as a fundamental optical resonance at a free-space wavelength near 1,550 nm, with a theoretical vacuum coupling rate of $g_0/2\pi = 860$ kHz. All measurements presented here are performed at room temperature and pressure. The thermal Brownian motion of the acoustic resonance manifests as a Lorentzian response centred around ω_m in the noise power spectral density (NPSD), $S(\omega)$, of the cavity reflection photocurrent, as shown in Fig. 1b. The linewidth of this Lorentzian is $\gamma = \gamma_i + \gamma_{OM}$, where γ_i is the intrinsic acoustic energy damping rate and $\gamma_{OM} = \pm 4G^2/\kappa$ is the optomechanically induced damping rate due to dynamical back action when pumping on the red or blue sideband, respectively[21]. By measuring this linewidth as a function of n_c for both red and blue detuning, we extract $\gamma_i/2\pi = 3$ MHz and $g_0/2\pi = 645$ kHz. The optical cavity reflection spectrum shown in Fig. 1c reveals a total optical energy decay rate of $\kappa/2\pi = 817$ MHz and a decay rate into the detection channel of $\kappa_e/2\pi = 425$ MHz.

To determine the feasibility of the phonon counting scheme for measurements of the OMC in the quantum regime, one must determine the signal-to-noise ratio (SNR). While the signal level depends on the above-measured optomechanical cavity parameters as well as the overall efficiency of photon detection, any detection events that do not correspond to Raman-scattered sideband photons will contribute as noise. These noise counts originate in dark counts of the single-photon detectors (SPDs) and photodetection of light at frequencies other than ω_c, most notably in the unscattered portion of the laser pump. Two
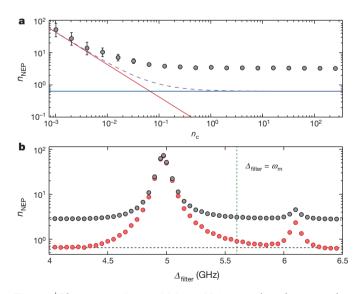
cascaded tunable, commercially available Fabry–Perot filters (Micron Optics, FFP-TF2), with bandwidths of 50 MHz and free spectral ranges of 20 GHz, are used to attenuate the laser pump by a factor $A > 80$ dB relative to the peak transmission at ω_c (see Fig. 1d). The transmission of the filters is then detected in the HBT apparatus by WSi-based superconducting nanowire SPDs[23] operating at a system detection efficiency of $\eta_{SPD} \approx 70\%$ and a dark count rate of $\Gamma_{dark} = 4$ Hz.

A useful parameterization of these quantities is the amount of noise (in units of mechanical occupation quanta) as a proportion of the signal generated by a single phonon in the OMC. Alternatively, this noise-equivalent phonon number n_{NEP} can be interpreted as the mechanical occupation which would produce an SNR of 1. We obtain n_{NEP} then by dividing the noise count rates by the per-phonon sideband photon count rate $\Gamma_{SB,0} = \eta|\gamma_{OM}|$, where η is the total efficiency of the set-up, including the system efficiency of the SPDs as well as optical insertion loss along the path from cavity to detector. For a coherent pump, this yields:

$$n_{NEP} = \frac{\kappa^2\Gamma_{dark}}{4\eta\kappa_e g_0^2 n_c} + A\left(\frac{\kappa\omega_m}{2\kappa_e g_0}\right)^2 \qquad (1)$$

The above equation makes clear the benefits of large cavity-enhanced optomechanical coupling g_0, both in terms of the low power sensitivity limited by detector dark counts and the high power sensitivity limited by pump bleed-through. Further details about the detection set-up, optical and mechanical spectroscopy, and derivation of n_{NEP} can be found in the Supplementary Information.
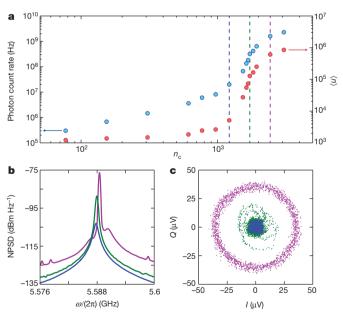
In Fig. 2a we show the measured sensitivity of the phonon counting set-up for $\Delta = \omega_m$ (filled grey circles), as well as the expected theoretical curve (dashed line) given by equation (1). The noise count rate is measured with the pump beam detuned far off-resonance from the optical cavity, $\Delta \gg \omega_m$, which eliminates the signal due to motional sideband photons but does not change the reflected pump signal or SPD dark counts. To determine $\Gamma_{SB,0}$, the sideband count rate is measured at low n_c where back action is negligible and the mechanical

Figure 2 | Phonon counting sensitivity. a, Noise equivalent phonon number n_{NEP} versus intracavity photon number n_c calculated using the measured signal and noise count rates for our current set-up (filled grey circles). Solid lines indicate the theoretically expected contributions due to dark counts (red) and pump bleed-through (blue), based on the measured system efficiency and pump suppression, with the sum of the two contributions displayed as a purple dashed line. Error bars show one s.d. determined from the measured count rates, assuming Poissonian counting statistics. **b**, n_{NEP} versus filter–pump detuning Δ_{filter} for $n_c \approx 65$, with (red) and without (grey) an additional C-band band-pass filter inserted. The vertical green line indicates the detuning corresponding to the data from **a**, and the horizontal black line indicates the expected limiting sensitivity.

mode occupancy $\langle n \rangle$ is equal to the room temperature thermal bath occupancy of $n_b \approx 1{,}100$. Since $\Gamma_{SB,0}$ scales linearly with n_c, we can determine $\Gamma_{SB,0}$ for all n_c from this single measurement without relying on calibration of the optomechanical back action. We can then compute n_{NEP} by dividing this noise count rate by $\Gamma_{SB,0}$ at each value of n_c. The measured sensitivity follows the expected curve at low power due to detector dark counts (solid red curve), but at high n_c saturates to a value several times larger than expected for the filter suppression of the pump (solid blue curve). In order to better understand this excess noise, Fig. 2b shows measurements of the n_{NEP} as a function of filter–pump detuning, Δ_{filter}, at a high power where the pump transmission dominates the noise ($n_c \approx 65$). A strong dependence on Δ_{filter} is observed, with a peak in the noise at 5 GHz and a secondary peak at 6.1 GHz, consistent with the phase noise of our pump laser[21]. With the addition of a C-band bandpass filter before the SPD to remove broadband spontaneous emission from the pump laser, and at frequencies far from the laser phase-noise peaks, the measured n_{NEP} agrees well with the theoretical predictions based on the filter pump suppression (horizontal dashed curve). At the relevant detuning of $\Delta_{filter} = \omega_m$ (vertical dashed curve), we measure a limiting sensitivity of $n_{NEP} = 0.89 \pm 0.05$. While this sensitivity is directly measured at $n_c \approx 65$, n_{NEP} is observed in Fig. 2a to be pump-limited for $n_c \gtrsim 1$, implying that our current set-up achieves $n_{NEP} < 1$ for n_c of order unity.
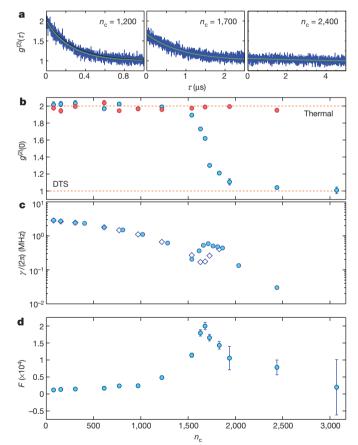
In what follows, we focus on measurements made with a blue-detuned pump ($\Delta = -\omega_m$), in which the optomechanical back action results in instability and self-oscillation of the acoustic resonator[9,21]. The Stokes sideband count rate detected on a single SPD, shown versus n_c in Fig. 3a, displays a pronounced threshold, with an exponential increase in output power beginning at $n_c \approx 1{,}200$, where $C \equiv |\gamma_{OM}|/\gamma_i \approx 0.8$, in agreement with the expected onset of instability around $C = 1$ ($\gamma = 0$). This sharp oscillation threshold can also be observed from the measured noise power spectral density (NPSD; Fig. 3b), in which the amplitude of the mechanical spectrum is seen to rapidly

increase with a simultaneous reduction in linewidth, and in plots of the in-phase and in-quadrature components of the photocurrent fluctuations, which show a transition from thermal noise to a large-amplitude sinusoidal oscillation. Also shown in Fig. 3a is the inferred phonon occupancy $\langle n \rangle$. Below threshold, the photon count rate is related to $\langle n \rangle$ via the simple linear relation $\Gamma_{tot} = \eta |\gamma_{OM}| (\langle n \rangle + 1)$. At and above threshold, as detailed in the Supplementary Information, self-consistent determination of the oscillation amplitude indicates that even at our highest pump power the mechanical amplitude remains small enough that this linear approximation remains valid.

The statistical properties of the resonator near the self-oscillation threshold can also be characterized by measuring photon correlations using an HBT set-up as shown in Fig. 1. As noted earlier, blue-detuned pumping produces anti-normally ordered phonon correlations. In this case $g^{(2)}(\tau)$ refers to the anti-normally ordered second-order phonon correlation function, defined by $g^{(2)}(\tau) = \langle \hat{b}(0)\hat{b}(\tau)\hat{b}^\dagger(\tau)\hat{b}^\dagger(0) \rangle / \langle \hat{b}(0)\hat{b}^\dagger(0) \rangle^2$. For measurements made with a red-detuned pump, as shown in Fig. 4b, $g^{(2)}(\tau)$ refers to the normally ordered phonon correlation function, $g^{(2)}(\tau) = \langle \hat{b}^\dagger(0)\hat{b}^\dagger(\tau)\hat{b}(\tau)\hat{b}(0) \rangle / \langle \hat{b}^\dagger(0)\hat{b}(0) \rangle^2$. In the case of the classical states measured here, there is no observable difference between

Figure 3 | Phonon lasing. a, Phonon count rate (blue: left-hand vertical axis) and inferred phonon occupancy $\langle n \rangle$ (red: right-hand vertical axis) as a function of intracavity photon number n_c for $\Delta = -\omega_m$. Dashed lines indicate points below (blue), at (green) and above (magenta) threshold. **b**, Noise power spectral densities (NPSD) corresponding to the dashed lines in **a**. The small satellite peaks in the thermal emission background of the above-threshold spectrum correspond to beating of the phonon laser line with low-frequency modes of the nanobeam structure. **c**, Phase plots of the in-phase (I) and in-quadrature (Q) amplitudes of the optical heterodyne signal for each of the dashed lines in **a**, acquired in a 36 MHz span around 5.588 GHz over a 60 s time interval.

Figure 4 | Phonon intensity correlations. a, Normalized anti-normally ordered second-order intensity correlation function $g^{(2)}(\tau)$ for $\Delta = -\omega_m$, shown below, at and above threshold (left, middle and right, respectively). Green lines show a simple exponential fit, while black lines indicate the expected theoretical curve using decay rates measured from fitting the NPSD linewidth. **b**, Phonon correlation at zero time delay versus n_c for $\Delta = -\omega_m$ (blue) and $\Delta = \omega_m$ (red). The top and bottom dashed lines indicate the expected values for purely thermal or displaced thermal states (DTS), respectively. Error bars show one s.d. determined from the fit value of $g^{(2)}(0)$. **c**, Mechanical decay rate versus n_c for $\Delta = -\omega_m$, determined from the measured linewidth of the NPSD (circles) and from the exponential fit to $g^{(2)}(\tau)$ (diamonds). **d**, Fano factor versus n_c. Error bars show one s.d. determined from the measured count rates, assuming Poissonian counting statistics, and the fit value of $g^{(2)}(0)$.

the normally and anti-normally ordered correlation functions. As the oscillation threshold is crossed, the state of the acoustic resonator will transition from a thermal state into a displaced thermal state (DTS), and the normalized phonon intensity correlation function near $\tau = 0$ should show a transition from bunching ($g^{(2)}(0) > 1$) to Poissonian statistics ($g^{(2)}(\tau) = 1$ for all τ). Plots of $g^{(2)}(\tau)$ below, at, and above threshold are shown in Fig. 4a. Below threshold, bunching is clearly visible, with $g^{(2)}(0) = 2$ as expected for a purely thermal state. In Fig. 4b $g^{(2)}(0)$ is plotted versus n_c for both blue- and red-detuned pump light. For a blue-detuned pump, a smooth decrease from $g^{(2)}(0) = 2$ to $g^{(2)}(0) = 1$ is observed in the threshold region, while for a red-detuned pump, the oscillator is observed to remain in a thermal state through threshold and beyond. The decay rate of the acoustic resonator, measured from both the linewidth of the NPSD and from an exponential fit to $g^{(2)}(\tau)$ below threshold, is plotted in Fig. 4c. The decay rate as measured from the NPSD, which includes both phase and amplitude fluctuations, is seen to increase around threshold before continuing to decrease. This behaviour is commonly observed in semiconductor lasers where a coupling exists between the gain and the cavity refractive index, and a similar effect arises in optomechanical oscillators due to the optical spring effect[24]. The decay rate measured from $g^{(2)}(\tau)$, on the other hand, which measures intensity fluctuations, begins to deviate from the measured linewidth in the vicinity of threshold. Thermal phonon emission dictates a strict correspondence between the second-order and first-order coherence functions[4]; however, above threshold where the phonon statistics are no longer purely thermal, such a deviation is possible, and in fact predicted for self-sustaining oscillators[25]. The Fano factor, defined as $F = (\Delta n)^2/\langle n \rangle = 1 + \langle n \rangle (g^{(2)}(0) - 1)$, provides additional statistical information about the fluctuations of the oscillator, and is useful for defining a precise oscillator threshold[26] as well as distinguishing between states that may have similar or identical values of $g^{(2)}(0)$ (for example, a coherent state versus a DTS)[24]. The Fano factor of our mechanical oscillator, computed from the measured $g^{(2)}(0)$ and the inferred values of $\langle n \rangle$, is displayed in Fig. 4d and shows the expected increase and peak in fluctuations at threshold. Above threshold, the Fano factor drops again due to saturation in the optomechanical gain, approaching a measured value consistent with that expected for a DTS ($F \approx 2n_b + 1$).

Although we have emphasized the analogy between the optomechanical oscillator studied here and a laser, there are unique differences which arise owing to the intrinsically nonlinear nature of the radiation pressure interaction in an optomechanical cavity. Recent theoretical studies[15,24,27] indicate that a laser-driven optomechanical oscillator will enter a non-classical mechanical state with anti-bunched phonon statistics ($F < 1$), and under slightly more restrictive conditions, strongly negative Wigner density. Surprisingly, this is predicted to be observable even for classical parameters, that is, outside the single-photon strong-coupling regime ($g_0/\kappa < 1$), and in the presence of thermal noise. Beyond phonon correlation spectroscopy of optomechanical oscillators, it is envisioned that sensitive photon counting of the filtered motional sidebands may be used in the preparation and heralding of non-Gaussian quantum states of a mechanical resonator[10]. For the OMC cavities of this work, with their large optomechanical coupling rate and near millisecond-long thermal decoherence time at temperatures less than 1 K (ref. 28), the phonon addition and subtraction processes of ref. 10 should be realizable with high fidelity and at rates approaching a megahertz. Whether for studies of the quantum behaviour of mesoscopic mechanical objects or in the context of proposed quantum information processing architectures using phonons and photons[8], such photon counting methods are an attractive way of introducing a quantum nonlinearity into the cavity optomechanical system.

1. Hoover, E. E. & Squier, J. A. Advances in multiphoton microscopy technology. *Nature Photon.* **7**, 93–101 (2013).
2. Hadfield, R. H. Single-photon detectors for optical quantum information applications. *Nature Photon.* **3**, 696–705 (2009).
3. Hanbury Brown, R. & Twiss, R. Q. A test of a new type of stellar interferometer on Sirius. *Nature* **178**, 1046–1048 (1956).
4. Glauber, R. J. The quantum theory of optical coherence. *Phys. Rev.* **130**, 2529–2539 (1963).
5. Aspelmeyer, M., Kippenberg, T. J. & Marquardt, F. Cavity optomechanics. *Rev. Mod. Phys.* **86**, 1391–1452 (2014).
6. Suh, J. *et al.* Mechanically detecting and avoiding the quantum fluctuations of a microwave field. *Science* **344**, 1262–1265 (2014).
7. Stannigel, K., Rabl, P., Sørensen, A. S., Lukin, M. D. & Zoller, P. Optomechanical transducers for quantum-information processing. *Phys. Rev. A* **84**, 042341 (2011).
8. Stannigel, K. *et al.* Optomechanical quantum information processing with photons and phonons. *Phys. Rev. Lett.* **109**, 013603 (2012).
9. Grudinin, I. S., Lee, H., Painter, O. & Vahala, K. J. Phonon laser action in a tunable two-level system. *Phys. Rev. Lett.* **104**, 083901 (2010).
10. Vanner, M. R., Aspelmeyer, M. & Kim, M. S. Quantum state orthogonalization and a toolset for quantum optomechanical phonon control. *Phys. Rev. Lett.* **110**, 010504 (2013).
11. Galland, C., Sangouard, N., Piro, N., Gisin, N. & Kippenberg, T. J. Heralded single-phonon preparation, storage, and readout in cavity optomechanics. *Phys. Rev. Lett.* **112**, 143602 (2014).
12. Børkje, K., Nunnenkamp, A. & Girvin, S. M. Proposal for entangling remote micromechanical oscillators via optical measurements. *Phys. Rev. Lett.* **107**, 123601 (2011).
13. Lee, K. C. *et al.* Entangling macroscopic diamonds at room temperature. *Science* **334**, 1253–1256 (2011).
14. Clerk, A. A., Devoret, M. H., Girvin, S. M., Marquardt, F. & Schoelkopf, R. J. Introduction to quantum noise, measurement, and amplification. *Rev. Mod. Phys.* **82**, 1155–1208 (2010).
15. Qian, J., Clerk, A. A., Hammerer, K. & Marquardt, F. Quantum signatures of the optomechanical instability. *Phys. Rev. Lett.* **109**, 253601 (2012).
16. Kronwald, A., Ludwig, M. & Marquardt, F. Full photon statistics of a light beam transmitted through an optomechanical system. *Phys. Rev. A* **87**, 013847 (2013).
17. Kimble, H. J., Dagenais, M. & Mandel, L. Photon antibunching in resonance fluorescence. *Phys. Rev. Lett.* **39**, 691–695 (1977).
18. Pike, R. 50th anniversary of the laser. *J. Eur. Opt. Soc. Rapid Publ.* **5**, 10047s (2010).
19. Chan, J. *et al.* Laser cooling of a nanomechanical oscillator into its quantum ground state. *Nature* **478**, 89–92 (2011).
20. Chan, J., Safavi-Naeini, A. H., Hill, J. T., Meenehan, S. & Painter, O. Optimized optomechanical crystal cavity with acoustic radiation shield. *Appl. Phys. Lett.* **101**, 081115 (2012).
21. Safavi-Naeini, A. H. *et al.* Laser noise in cavity-optomechanical cooling and thermometry. *New J. Phys.* **15**, 035007 (2013).
22. Fetter, A. L. Intensity correlations in Raman scattering. *Phys. Rev.* **139**, A1616–A1623 (1965).
23. Marsili, F. *et al.* Detecting single infrared photons with 93% system efficiency. *Nature Photon.* **7**, 210–214 (2013).
24. Rodrigues, D. A. & Armour, A. D. Amplitude noise suppression in cavity-driven oscillations of a mechanical resonator. *Phys. Rev. Lett.* **104**, 053601 (2010).
25. Lax, M. Classical noise. V. Noise in self-sustained oscillators. *Phys. Rev.* **160**, 290–307 (1967).
26. Rice, P. R. & Carmichael, H. J. Photon statistics of a cavity-QED laser: a comment on the laser-phase-transition analogy. *Phys. Rev. A* **50**, 4318–4329 (1994).
27. Lörch, N., Qian, J., Clerk, A., Marquardt, F. & Hammerer, K. Laser theory for optomechanics: limit cycles in the quantum regime. *Phys. Rev. X* **4**, 011015 (2014).
28. Meenehan, S. M. *et al.* Silicon optomechanical crystal resonator at millikelvin temperatures. *Phys. Rev. A* **90**, 011803 (2014).

LETTER

In situ low–relief landscape formation as a result of river network disruption
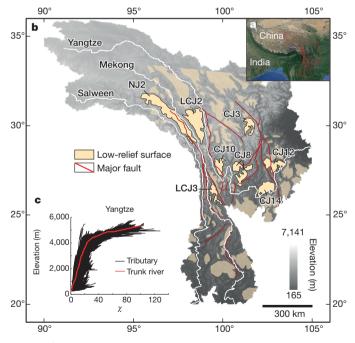
Rong Yang[1], Sean D. Willett[1] & Liran Goren[2]

Landscapes on Earth retain a record of the tectonic, environmental and climatic history under which they formed. Landscapes tend towards an equilibrium in which rivers attain a stable grade that balances the tectonic production of elevation and with hillslopes that attain a gradient steep enough to transport material to river channels. Equilibrium low-relief surfaces are typically found at low elevations, graded to sea level. However, there are many examples of high-elevation, low-relief surfaces, often referred to as relict landscapes[1,2], or as elevated peneplains[3]. These do not grade to sea level and are typically interpreted as uplifted old landscapes, preserving former, more moderate tectonic conditions[4]. Here we test this model of landscape evolution through digital topographic analysis of a set of purportedly relict landscapes on the southeastern margin of the Tibetan Plateau, one of the most geographically complex, climatically varied and biologically diverse regions of the world. We find that, in contrast to theory, the purported surfaces are not consistent with progressive establishment of a new, steeper, river grade, and therefore they cannot necessarily be interpreted as a remnant of an old, low relief surface. We propose an alternative model, supported by numerical experiments, in which tectonic deformation has disrupted the regional river network, leaving remnants of it isolated and starved of drainage area and thus unable to balance tectonic uplift. The implication is that the state of low relief with low erosion rate is developing *in situ*, rather than preserving past erosional conditions.

The Tibetan Plateau has uplifted over the past approximately 50 million years (Myr) and has been argued to be growing eastwards through a combination of block motion on strike-slip faults[5–7], uplift in response to lower crustal flow[8] and large-scale shortening in response to indentation of the Indian plate[9]. To the southeast of the plateau proper is a topographic transition, referred to as the 'Three Rivers' region, taking its name from the three large rivers (Salween, Mekong and Yangtze) that have incised the edge of the plateau, forming bedrock gorges up to 3 km deep (Fig. 1). Although surface expression of tectonic activity is not always evident, large strike-slip faults parallel the main rivers[10] and there is evidence for tectonic disturbance to the landscape, including captures of major rivers[11,12]. Geodynamic models suggest that the region, situated near the corner of the indenting India plate, has been subjected to large horizontal strains[9,13]. Thermochronometry suggests that the major rivers accelerated their incision about 10 Myr ago (Ma), although paleo-altimetry based on stable isotope data suggests that most of the area has been near its current elevation for more than 40 Myr[14,15]. A regional low-relief surface, or relict landscape, with low erosion rates[16–19] has been identified in the headwaters of the three rivers and along the interfluves in the middle and lower reaches, and its existence has been used to argue for relatively recent uplift[1,20].

The transient response of a landscape to a change in uplift rate is dominated by changes in river channel slope. Channel slope has a first-order dependence on discharge or its proxy, drainage area, so

measures of channel slope must be normalized for drainage area. Similar to past treatments of the problem[21–23], we transform distance from base level, x, along the channel to the surrogate quantity, χ:

$$\chi(x) = \int_0^x \left(\frac{P_0 A_0}{A(x')P(x')} \right)^{\frac{m}{n}} \mathrm{d}x' \quad (1)$$

where m and n are empirical, non-integer constants, P is precipitation rate and A is the upstream drainage area[24]. P_0 and A_0 are arbitrary scaling factors for the precipitation rate and drainage area, respectively. The change in channel elevation with respect to χ is referred to as channel steepness and provides a measure of the local erosion rate. For

Figure 1 | The study area of the 'Three Rivers'. a, The catchment boundaries of the Salween, Mekong and Yangtze Rivers in southeast Asia. **b**, Map of low-relief areas within the 'Three Rivers' region. Trunk channel of the Salween, Mekong and Yangtze Rivers shown in white. Major faults shown in red. Yellow shading indicates low-relief surfaces as identified in ref. 1. Surfaces singled out for our study and shown in subsequent figures are labelled. **c**, Normalized elevation plot (χ-plot) for the Yangtze. Red curve is the profile of the main stem. χ is calculated from equation (1) using $m/n = 0.45$. Precipitation rate is shown in Extended Data Fig. 1. The comparable information for the Salween and Mekong is shown in Extended Data Fig. 2. The same data for a range of values of m/n are shown in Extended Data Fig. 2. Erosion rate scales with the slope of the plot[22], so the scatter in the slope of the tributaries implies variability in the erosion rate[17].

[1]Geological Institute, Swiss Federal Institute of Technology, 8092 Zurich, Switzerland. [2]Geological and Environmental Sciences, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel.

a steady, constant uplift rate, erosion rate equals the uplift rate and the channel steepness, K_s, is proportional to the uplift rate, U, as
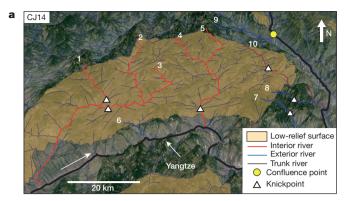
$$K_s = \left(\frac{U}{KA_0^m}\right)^{\frac{1}{n}} \quad (2)$$

where K reflects the basin runoff characteristics and rock erodibility[25]. According to prevailing theory, a channel steepness transient initiates at base level and propagates upstream as a kinematic wave[25–27]. By normalizing for drainage area, we can compare rivers of variable size across a landscape given that channels with a common uplift pattern and history should collapse onto a single profile in elevation–χ space[22]. Even with added complexities in the incision law[28], area–slope scaling at steady state remains a robust feature of most landscapes, so that a χ-plot gives a valuable metric for the state of disequilibrium. The χ-plot of the Yangtze River above the Sichuan Basin using the precipitation data from the Tropical Rainfall Measurement Mission (Extended Data Fig. 1) shows high steepness for χ less than about 20 and much lower steepness for larger χ (Fig. 1c), reflecting the low relief of the Tibetan plateau. However, scatter in this plot is very large at all elevations. The Salween and Mekong Rivers exhibit similarly large scatter (Extended Data Fig. 2). This variability in steepness is consistent with the existence of low-steepness landscape fragments throughout the region, but it is not consistent with the idea of a common history of uplift. However, over a region as large as the 'Three Rivers', we expect temporal and spatial variations in uplift and rock erodibility, so the analysis is best applied at a smaller scale.

To test whether there is a common uplift history for individual low-relief surfaces, we generated χ-plots for the rivers draining the interior of individual 'relict' surfaces as identified in ref. 1. These profiles were compared with profiles of the rivers draining the exterior of the low-relief landscape and with the main trunk of the nearest of the three rivers. We attempted to select low-relief landscapes interior to one of the three major drainage basins, and only considered tributary rivers having a common confluence close to the low-relief surface. Any differences between channel profiles must arise above their common confluence, confirming that we were looking at local processes. In total, we investigated eight regions (Fig. 2 and Extended Data Figs 3–9).

All regions showed remarkably consistent river profile characteristics (see, for example, Fig. 2). It is clear that there is no common uplift history that can explain the aggregate set of rivers, even in the limited area of a single low-relief surface, and independent of the selection of exponents in equation (1) (Extended Data Fig. 2). The kinematic wave model of channel equilibration predicts collapse of all rivers onto a single elevation–χ profile[22], which is not observed. All tributaries mapped in Fig. 2 deviate in steepness, as represented by the slope of the χ-plot, upstream from their confluence with the main stem of the Yangtze River. What is most remarkable is that they deviate systematically. All rivers draining the interior of the low-relief surface plot below the regional average; all rivers draining the exterior perimeter of the surface have a steeper slope and plot above the interior-draining rivers and above the trunk river. In every case, there is a specific range of χ over which the interior river channels have a lower steepness than the exterior channels. In addition, the large contrast in χ across water divides surrounding the low-relief catchments suggests that the regional network does not have a stable geometry and thus does not retain an old, near-equilibrium state[21]. Differences in steepness can arise from differences in rock erodibility, uplift rate or precipitation rate, and individual features may be attributed to these factors; however, given the restricted area of each of these examples, and the individual geometry of each low-relief surface, we can exclude these factors as a general explanation for the pattern of river profile diversity in all eight areas.

The other potential source of variation in the χ-plots arises from changes in drainage basin area by river capture and divide migration.
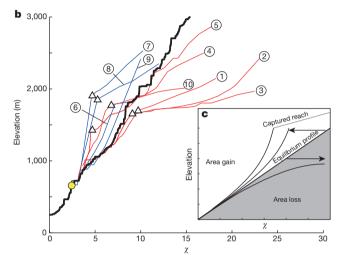
Figure 2 | River courses and χ-plots for region of low-relief, 'relict' landscape CJ14 in the Yangtze drainage. See Fig. 1 for location. **a**, Perspective view of the landscape. The low-relief surface is shaded and the courses of ten rivers are keyed by colour and number to the χ-plots in **b**. **b**, Graph of χ-plots for rivers shown in Fig. 2a. Red tributaries are primarily interior to the low-relief landscape; blue rivers drain the exterior. Yellow point in each frame shows downstream point common to all rivers. Triangles indicate prominent inflections in profiles and their locations in Fig. 2a. **c**, Schematic of response of χ-plot to instantaneous area change or discrete river capture. Loss of drainage area shifts plot to the right; gain of area shifts plot to the left. A discrete capture event may preserve the slope of the captured reach. Note that interior rivers (1, 2, 3) are all shifted into the area-loss field, although the lower reaches may have equilibrated with the main stem of the Yangtze. Tributaries 4–10 are shifted into the area-gain field. Tributaries 4, 5, 7, 8 and 10 appear to have recently captured additional area from the upper surface, thus exhibiting a kinked profile. Examples of this analysis applied to other low-relief landscapes are shown in Extended Data Figs 3–9.

Temporal changes in drainage area produce distinctive changes in the χ-profiles[21]. Given that χ depends inversely on drainage area, a loss of area pushes an equilibrium profile towards higher χ, preferentially affecting the upper reaches of a river, as shown in Fig. 2c. Conversely, an increase in drainage area decreases χ, pulling the profile to the left in a χ-plot. In the case of a large river capture, the river below the capture point is pulled to the left and its plot steepened, with the captured reach retaining its original slope, thus giving a distinctive, kinked plot (Fig. 2c). In all cases, area change leads to a disequilibrium state of the river channel, which will evolve back towards equilibrium through changes in erosion rate and channel steepness. There is an important positive feedback to this process, in that, during the disequilibrium phase, area loss leads to a lower erosion rate and thereby to surface uplift and higher elevation. Higher elevation increases vulnerability to further area loss by either capture or divide migration. A river that is the victim of drainage-area loss is therefore characterized by high
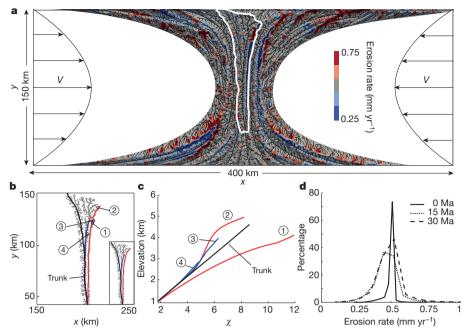
Figure 3 | Numerical model of landscape evolution governed by stream-power river incision in response to tectonic uplift and horizontal strain using the Divide and Capture (DAC) landscape evolution model[29]. **a**, River network configuration and erosion rate shown at model time of 25.8 Myr. Tectonic rock uplift rate is constant in space and time at 0.5 mm yr^{-1}. Initial topography is near steady state with the same uplift rate and no horizontal motion. All boundaries are fixed elevation at nominal sea level. At time zero, domain is subjected to horizontal contractive field with a double parabolic form and a maximum velocity of 15 mm yr^{-1} (see Methods). Shortening vectors shown as black arrows. Only lower (south) half of the symmetrical model is shown, so upper (north) edge of figure is approximately the main divide of an approximately symmetric domain. Full model is shown in the Supplementary Video. **b**, A select drainage basin (see location highlighted in white in **a**) exhibiting typical erosion rate variations. The red segment of river 2 has been captured from river 1. The inset shows the drainage basin configuration just before river capture. **c**, Graph of χ-plot of river basin shown in **b**. River steepness (slope of χ-plot) is low in tributaries that have recently lost area and high in tributaries that have gained area. Note similarities in overall basin structure with natural example in Fig. 2. **d**, Distribution of erosion rate across full model domain at three times, showing increase in erosion rate variance in response to tectonic strain. Note decrease in mean and negative skewness in distribution, reflecting the perseverance of low-erosion rate regions.

elevation, low channel steepness, low erosion rates and is typically surrounded by rivers that are aggressively advancing into its current drainage basin[21]. These are precisely the characteristics of the regional landscapes portrayed here (Fig. 2 and Extended Data Figs 3–9).

To demonstrate this mechanism and to show that tectonic shortening and shear deformation does produce such a landscape, we constructed a numerical model of a landscape experiencing a non-uniform shortening rate representing the indentation of the Indian plate into southeast Asia[9,13]. The model solves for stream-power incision of river channels in a river network, subjected to tectonic uplift and horizontal motion, and includes an explicit description of the water divide motion[29] (Methods). In our numerical experiment (Fig. 3 and Supplementary Video), we impose a pincer-type horizontal strain-rate field on an initially steady state topography, but retain the initial rock uplift rate. The tectonic strain deforms the river basins and induces multiple river captures as the channel network attempts to regain a more stable configuration. These changes in network configuration also induce variations in river steepness (Fig. 2c) with a corresponding increase in variance of erosion rate (equation (2)). In particular, a large fraction of the model domain has erosion rates lower or higher than the imposed tectonic uplift rate (Fig. 3d). Low erosion rates are in part due to the lengthening of flow paths by extensional strain, but the lowest erosion rate regions are typically victims of a large river capture. Given the area-loss feedback, these low-erosion-rate regions are subject to multiple subsequent captures, as well as continuous inward divide migration by steeper neighbouring basins. These processes prolong the transient phase of landscape response and result in isolated catchments with anomalously high elevations and low erosion rates, much as is observed in the 'Three Rivers' region (compare Fig. 3c with Fig. 2b or Extended Data Figs 3–9).

Our numerical model includes several simplifications. The imposed strain field is continuous, whereas in nature surface deformation is commonly brittle, with localized faults. The use of a continuous strain field has the tendency to minimize channel network disruption; the same strain distributed across discrete faults would probably have a much larger effect on the channel network geometry. We also do not include sediment production and deposition in our model. A river that loses upstream area loses sediment carrying capacity and may not be able to transport sediment supplied by local hillslopes. This sediment would be locally deposited, preventing additional channel erosion[28]. Thus our model provides a minimum estimate of the variations in erosion rate expected through channel network disruption.

Drainage basin disruption slows, but does not prevent the re-equilibration of low erosion rate catchments to base level; in fact, we observe steepening of the lower reaches of many rivers with convex inflections separating steep and gentle river segments (Fig. 2 and Extended Data Figs 3–9). However, the height and position of these inflections are typically unique to an individual drainage basin, inconsistent with the idea of a regional wave of incision propagating upstream. As an end-member interpretation, each low-relief catchment could correspond to a specific large capture event. In some cases, it is possible to determine this. For example, surface CJ12 (Extended Data Fig. 8) is downstream from the former course of the Dadu River[12]. However, in most cases, low-relief surfaces result not from a single event, but through a series of captures accompanied by continuous divide migration, making it difficult to assign attribution to a single event. We also cannot exclude the possibility that much of the region has experienced an acceleration in rock uplift. In fact, an increase in uplift rate would be expected to accompany the shortening of our simulation. In addition, the upper 'Three Rivers' region drains

the Tibetan Plateau, whose existence over a much larger area suggests that other tectonic or climatic processes are important to formation of the plateau landscape that includes the uppermost low-relief surfaces[30]. However, our model suggests that an acceleration in uplift rate is not required to form most of the landscape of the 'Three Rivers' region, nor is an initially low-relief landscape required. We suggest, rather, that the 'relict' landscapes are forming and being modified, *in situ* through loss of drainage area. Given the positive feedback inherent to drainage area loss and gain, it is unlikely that these, or any high, isolated surface will retain an unaltered record of past geomorphic conditions or elevations.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Clark, M. K. *et al.* Use of a regional, relict landscape to measure vertical deformation of the eastern Tibetan Plateau. *J. Geophys. Res.* **111**, http://dx.doi.org/10.1029/2005JF000294 (2006).
2. Schoenbohm, L. M., Whipple, K. X., Burchfiel, B. C. & Chen, L. Geomorphic constraints on surface uplift, exhumation, and plateau growth in the Red River region, Yunnan Province, China. *Geol. Soc. Am. Bull.* **116**, 895–909 (2004).
3. Davis, W. Baselevel, grade and peneplain. *J. Geol.* **10**, 77–111 (1902).
4. Babault, J., Van Den Driessche, J., Bonnet, S., Castelltort, S. & Crave, A. Origin of the highly elevated Pyrenean peneplain. *Tectonics* **24**, TC2010 (2005).
5. Tapponnier, P., Peltzer, G., Le Dain, A., Armijo, R. & Cobbold, P. Propagating extrusion tectonics in Asia: new insights from simple experiments with plasticine. *Geology* **10**, 611–616 (1982).
6. England, P. & Molnar, P. Right-lateral shear and rotation as the explanation for strike-slip faulting in eastern Tibet. *Nature* **344**, 140–142 (1990).
7. Tapponnier, P. Oblique stepwise rise and growth of the Tibet Plateau. *Science* **294**, 1671–1677 (2001).
8. Clark, M. K. & Royden, L. H. Topographic ooze: building the eastern margin of Tibet by lower crustal flow. *Geology* **28**, 703–706 (2000).
9. Hallet, B. & Molnar, P. Distorted drainage basins as markers of crustal strain east of the Himalaya. *J. Geophys. Res. Solid Earth* **106**, 13697–13709 (2001).
10. Burchfiel, B. C. & Chen, Z. *Tectonics of the Southeastern Tibetan Plateau and its Adjacent Foreland* 77–164 (Geological Society of America, 2012).
11. Clift, P. D., Blusztajn, J. & Nguyen, A. D. Large-scale drainage capture and surface uplift in eastern Tibet–SW China before 24 Ma inferred from sediments of the Hanoi Basin, Vietnam. *Geophys. Res. Lett.* **33**, http://dx.doi.org/10.1029/2006GL027772 (2006).
12. Clark, M. K. *et al.* Surface uplift, tectonics, and erosion of eastern Tibet from large-scale drainage patterns. *Tectonics* **23**, http://dx.doi.org/10.1029/2002TC001402 (2004).
13. England, P. & Houseman, G. Finite strain calculations of continental deformation. 2. Comparison with the India-Asia Collision Zone. *J. Geophys. Res.* **91**, 3664–3676 (1986).
14. Rowley, D. B. & Currie, B. S. Palaeo-altimetry of the late Eocene to Miocene Lunpola basin, central Tibet. *Nature* **439**, 677–681 (2006).
15. Hoke, G. D., Liu-Zeng, J., Hren, M. T., Wissink, G. K. & Garzione, C. N. Stable isotopes reveal high southeast Tibetan Plateau margin since the Paleogene. *Earth Planet. Sci. Lett.* **394**, 270–278 (2014).
16. Henck, A. C., Huntington, K. W., Stone, J. O., Montgomery, D. R. & Hallet, B. Spatial controls on erosion in the Three Rivers region, southeastern Tibet and southwestern China. *Earth Planet. Sci. Lett.* **303**, 71–83 (2011).
17. Ouimet, W. B., Whipple, K. X. & Granger, D. E. Beyond threshold hillslopes: Channel adjustment to base-level fall in tectonically active mountain ranges. *Geology* **37**, 579–582 (2009).
18. Duvall, A. R., Clark, M. K., Avdeev, B., Farley, K. A. & Chen, Z. Widespread late Cenozoic increase in erosion rates across the interior of eastern Tibet constrained by detrital low-temperature thermochronometry. *Tectonics* **31**, TC3014 (2012).
19. Wilson, C. J. L. & Fowler, A. P. Denudational response to surface uplift in east Tibet: evidence from apatite fission-track thermochronology. *Geol. Soc. Am. Bull.* **123**, 1966–1987 (2011).
20. Clark, M. K. *et al.* Late Cenozoic uplift of southeastern Tibet. *Geology* **33**, 525–528 (2005).
21. Willett, S. D., McCoy, S. W., Perron, J. T., Goren, L. & Chen, C. Y. Dynamic reorganization of river basins. *Science* **343**, http://dx.doi.org/10.1126/science.1248765.
22. Perron, J. T. & Royden, L. An integral approach to bedrock river profile analysis. *Earth Surf. Process. Landf.* **38**, 570–576 (2012).
23. Royden, L. & Taylor Perron, J. Solutions of the stream power equation and application to the evolution of river longitudinal profiles. *J. Geophys. Res. Earth Surf.* **118**, 497–518 (2013).
24. Ferrier, K. L., Huppert, K. L. & Perron, J. T. Climatic control of bedrock river incision. *Nature* **496**, 206–209 (2013).
25. Whipple, K. X. & Tucker, G. E. Dynamics of the stream-power river incision model: Implications for height limits of mountain ranges, landscape response timescales, and research needs. *J. Geophys. Res. Solid Earth* **104**, 17661–17674 (1999).
26. Seidl, M. & Dietrich, W. The problem of channel erosion into bedrock. *Catena, Suppl.* **23**, 101–124 (1992).
27. Weissel, J. & Seidl, M. in *Rivers over Rock: Fluvial Processes in Bedrock Channels* (American Geophysical Union Geophysical Monograph Series Vol. 107) (eds Tinkler, K. J. & Wohl, E. E.) 189–206 (American Geophysical Union, 1998).
28. Sklar, L. & Dietrich, W. E. in *Rivers over Rock: Fluvial Processes in Bedrock Channels* (American Geophysical Union Geophysical Monograph Series Vol. 107) (eds Tinkler, K. J. & Wohl, E. E.) 237–260 (American Geophysical Union, 1998).
29. Goren, L., Willett, S. D., Herman, F. & Braun, J. Coupled numerical-analytical approach to landscape evolution modeling. *Earth Surf. Process. Landf.* **39**, 522–545 (2014).
30. Liu-Zeng, J., Tapponnier, P., Gaudemer, Y. & Ding, L. Quantifying landscape differences across the Tibetan plateau: implications for topographic relief evolution. *J. Geophys. Res.* **113**, http://dx.doi.org/10.1029/2007JF000897 (2008).

METHODS

Topographic analysis was done using the SRTM3 digital elevation model, which has a resolution of approximately 90 m (ref. 31). A threshold drainage area of $5\,km^2$ was used to exclude regions dominated by debris flows or hillslope processes. We extracted all information such as longitude, latitude, elevation, flow direction, flow length, stream order and flow accumulation from the filled digital elevation model. The averaged annual precipitation rates were obtained with a spatial resolution of approximately $5\,km \times 5\,km$ based on the 12-year Tropical Rainfall Measurement Mission data during the period 1998–2009 following the methods described in ref. 32. The χ-value for each pixel was calculated according to equation (1), using the Tropical Rainfall Measurement Mission precipitation data. Note that the equation (2) definition of χ corresponds to a version of χ' in ref. 21. Calculations were done assuming concavity $m/n = 0.45$, a scaling area A_0 of $1\,m^2$ and a scaling precipitation rate P_0 of $1\,m\,yr^{-1}$. This was selected according to the methodology in ref. 21. Other values of m/n are shown in Extended Data Fig. 2. Use of another concavity changes the position of the trunk river relative to the tributaries, but does not change the relative position or steepness of the interior- and exterior-draining rivers. To reduce the number of pixels for calculation, we applied a skipping factor; the data were resampled every five pixels for the Salween and Mekong and every ten pixels for the Yangtze but not for those pixels at confluences to maintain the river network structure. For the trunk river profile in Figs 2 and Extended Data Figs 3–9, flats due to filling were removed.

The landscape evolution simulation was generated using the coupled numerical–analytical model DAC (Divide and Capture)[29]. This model implements the physics of water divide motion and stream capture. We defined a rectangular domain size of $400\,km \times 300\,km$. We used $n = 1$ for the slope exponent, $m = 0.5$ for the area exponent and $K = 1.0 \times 10^{-6}$ for the rock erodibility. The simulation was initialized by generating a steady-state topography from an initially random (Gaussian, uncorrelated, maximum elevation of $1\,m$) elevation field with a spatially uniform tectonic uplift rate of $0.5\,mm\,yr^{-1}$ and no horizontal motion. The base level on all four edges was fixed at a constant elevation, and precipitation rate and rock type were assumed to be steady and uniform. A symmetrical mountain range with maximum elevation of $5.4\,km$ was generated after 100 Ma. Using this steady-state topography as an initial condition, we imposed a horizontal velocity field to simulate the shortening in response to indentation of the corner of the Indian plate into Eurasia. Horizontal velocity has non-zero component only in the x direction (east–west) and varies parabolically in the y direction (north–south) from the boundary to the centre line of the model. Velocity varies linearly in x, from maximum values at the boundaries to zero along the centre axis. The midpoint axes in both dimensions are thus symmetry conditions for the velocity field. Upper and lower halves of the model can be regarded as independent realizations of the indentation problem. The simulation starts at 0 Ma and ends at 30 Ma with a time step length of 500 years.

Code availability: the landscape evolution model and parameters are available from any of the authors on request.

31. Jarvis, A., Reuter, H. I., Nelson, A. & Guevara, E. Hole-filled SRTM for the globe v. 4 (http://srtmcsi.cgiar.org) (2008).
32. Bookhagen, B. & Strecker, M. R. Orographic barriers, high-resolution TRMM rainfall, and relief variations along the eastern Andes. *Geophys. Res. Lett.* **35,** http://dx.doi.org/10.1029/2007GL032011 (2008).
33. Xu, G. & Kamp, P. J. Tectonics and denudation adjacent to the Xianshuihe Fault, eastern Tibetan Plateau: constraints from fission track thermochronology. *J. Geophys. Res. Solid Earth* **105,** 19231–19251 (2000).

LETTER

Tungsten isotopic evidence for disproportional late accretion to the Earth and Moon

Mathieu Touboul[1†], Igor S. Puchtel[1] & Richard J. Walker[1]

Characterization of the hafnium–tungsten systematics (^{182}Hf decaying to ^{182}W and emitting two electrons with a half-life of 8.9 million years) of the lunar mantle will enable better constraints on the timescale and processes involved in the currently accepted giant-impact theory for the formation and evolution of the Moon, and for testing the late-accretion hypothesis. Uniform, terrestrial-mantle-like W isotopic compositions have been reported[1,2] among crystallization products of the lunar magma ocean. These observations were interpreted to reflect formation of the Moon and crystallization of the lunar magma ocean after ^{182}Hf was no longer extant—that is, more than about 60 million years after the Solar System formed. Here we present W isotope data for three lunar samples that are more precise by a factor of $\geq$4 than those previously reported[1,2]. The new data reveal that the lunar mantle has a well-resolved ^{182}W excess of 20.6 ± 5.1 parts per million (±2 standard deviations), relative to the modern terrestrial mantle. The offset between the mantles of the Moon and the modern Earth is best explained by assuming that the W isotopic compositions of the two bodies were identical immediately following formation of the Moon, and that they then diverged as a result of disproportional late accretion to the Earth and Moon[3,4]. One implication of this model is that metal from the core of the Moon-forming impactor must have efficiently stripped the Earth's mantle of highly siderophile elements on its way to merge with the terrestrial core, requiring a substantial, but still poorly defined, level of metal–silicate equilibration.

Early applications of Hf–W isotopic system to lunar rocks were hampered by the effects of cosmic rays on ^{182}W, particularly due to production of ^{182}W from ^{181}Ta via neutron capture[5]. Consequently, more-recent studies have focused on the analysis of Ta-free metals extracted from lunar basalts and impact-melt rocks[1,6]. Here, we present W isotopic data for metals separated from two KREEP-rich Apollo 16 impact-melt rocks, obtained using new high-precision analytical methods[7] (KREEP indicates rocks rich in potassium (K), rare-earth elements (REE) and phosphorus (P)). In addition to W isotopic compositions and abundances, we determined abundances of Hf and the highly siderophile elements (HSE), as well as ^{187}Os/^{188}Os ratios, in order to assess the Hf/W ratio of the metal, examine the chemical signature of the impactor that produced the melt rocks, and evaluate potential contributions from meteoritic W. These impact-melt rocks were generated by basin-forming events, possibly during a period of late heavy bombardment about 3.9 Gyr ago[8]. The metal present in these rocks may have been derived from either the crustal target rocks, or the impactor that created the melt rocks. In either case, siderophile elements in the metals probably partially or wholly equilibrated with melt or vapour during the impact[9].

Metal separates from impact-melt rocks 68115,114, 68815,394 and 68815,396 (see Methods for nomenclature) have μ^{182}W values (where μ^{182}W is the deviation in p.p.m. of the ^{182}W/^{184}W ratio of the sample from that of the modern terrestrial mantle) of +23.3 ± 3.8 ($n = 3$, 2 s.d.), +18.1 ± 2.5 and +20.4 ± 2.9, respectively, which are identical within analytical uncertainty (Table 1). The new data are consistent with the previously published data for the same samples[1], but are considerably more precise (Fig. 1). Of greatest importance, the W isotopic compositions of the metals are now well resolved from the isotopic composition of the silicate portion of the modern Earth.

The positive W isotopic offset between the Moon and the silicate Earth can be attributed to one of several possible causes, including: (1) cosmogenic exposure effects, (2) contribution of W from the basin-forming impactor that created the melt rocks, (3) radiogenic ingrowth of ^{182}W in a high-Hf/W domain within the mantle of the Moon or of the giant impactor, or (4) disproportional late accretion to the Earth and Moon. These possibilities are considered below.

Cosmic ray exposure effects can be excluded as the cause of the isotopic offset. The Ta/Hf ratio of ~0.11 for KREEP[10], coupled with the measured Hf/W ratios of the metal separates (Table 1), gives a calculated Ta/W average of ~0.008 for the metals. Based on this ratio and the ~2 Myr exposure ages of the samples[11], a maximum effect of approximately −0.1 p.p.m. on the W isotopic compositions of the separated metals is estimated. This is well below our current level of analytical uncertainty, so no exposure corrections were applied, and the isotopic compositions of the metal separates are interpreted to have pre-exposure W isotopic compositions that are identical, within our long-term external analytical reproducibility.

Mass balance calculations indicate that the basin-forming impactor that led to the creation of the melt rocks contributed little W to the metals, and did not play a significant role in generating the isotopic offset. As a strongly incompatible element in silicate systems, W abundances are highly enriched in chemically evolved, KREEP-rich rocks[12], from which these impact-melt rocks were derived. Further, based on HSE abundances, basin-forming impactors with chondritic bulk compositions contributed no more than ~2–3% of their mass to impact-melt rocks[13]. Thus, if the concentrations of W in the Apollo 16 target rocks and the impactor are assumed to have been ~1,500 p.p.b., typical of KREEP-rich basalts[12], and ~180 p.p.b., typical of chondritic meteorites[14], respectively, a 3 wt% contribution of chondritic impactor mass to the melt rocks would have added only ~0.4% of the total W present in the impact-melt rock. If the impactor had a chondritic μ^{182}W value of −200, the indigenous lunar W isotopic composition would have been lowered by only ~0.8 p.p.m. (Extended Data Table 1).

The separated metals from 68815 are characterized by suprachondritic ^{187}Os/^{188}Os, Ru/Ir, Pt/Ir and Pd/Ir ratios (Table 2; Fig. 2), features that are common in chemically evolved iron meteorites[15]. Based on similar HSE characteristics in bulk samples of Apollo 16 impact-melt rocks, it has been argued that a chemically evolved iron meteorite, comparable to the group IVA iron meteorite Bushman Land, was involved in the generation of Apollo 16 impact-melt rocks[16]. Bushman Land is rich in siderophile elements, with 1,200 p.p.b. Ir and 460 p.p.b. W (ref. 15). Addition of no more than 0.8% of comparable material would be required to generate impact-melt rocks with the observed average Ir concentrations of <10 p.p.b. present in most Apollo 16 impact-melt rocks[16]. Incorporation of 1 wt% of an iron

[1]Isotope Geochemistry Laboratory, Department of Geology, University of Maryland, College Park, Maryland 20742, USA. †Present address: Laboratoire de Géologie de Lyon, Ecole Normale Supérieure de Lyon, Labex LIO, Université Lyon 1, 46 Allée d'Italie 69364 Lyon Cedex 7, France.

Table 1 | Tungsten isotopic compositions and W and Hf abundances of lunar metals

Samples	W (p.p.m.)	Hf (p.p.m.)	$\mu^{182}W$
68115,114 metal	32.7 ± 0.3	2.23 ± 0.05	+25.3 ± 4.6
			+21.5 ± 2.6
			+23.0 ± 1.7
68115,114 metal average (±2 s.d.)			+23.3 ± 3.8
68815,394 metal	22.8 ± 0.3	1.41 ± 0.02	+18.1 ± 2.5
68815,396 metal	36.3 ± 0.5	0.27 ± 0.01	+20.4 ± 2.9
Bulk lunar mantle (n = 3, ±2 s.d.)			+20.6 ± 5.1

meteorite with the same W concentration and with $\mu^{182}W = -330$, comparable to IVA irons[17], would have lowered the W isotopic compositions of the impact-melt rocks by only ~1 p.p.m., relative to the indigenous lunar signature (Extended Data Table 1).

As further evidence that the W isotopic compositions of the metal separates are nearly completely derived from the lunar target rocks, we note that the metals from the two pieces of 68815 are characterized by considerably different absolute and relative abundances of HSE, as well as $^{187}Os/^{188}Os$, yet their W isotopic compositions are identical within uncertainties. The differences in HSE are likely to reflect the incorporation of different proportions of HSE from two impactors into the two pieces. Incorporation of HSE from more than one impactor is common in lunar impact-melt rocks[18]. If our assumptions about the mass balance of W among target rocks and impactors are grossly incorrect, and significant but variable proportions of the W present in the metal separates were derived from different impactors, it is very likely they would also have different W isotopic compositions. This is not observed. We conclude that modifications to the indigenous lunar W isotopic composition by contamination from basin-forming impactors were minor, and that the average $\mu^{182}W$ value of +20.6 ± 5.1 (2 s.d.) for the three metal separates provides the current best estimate of the W isotopic composition of their parental mantle KREEP domain.

The observed isotopic offset between the Moon and the silicate Earth might also reflect *in situ* decay of ^{182}Hf in a high-Hf/W domain, formed as a consequence of the characteristics of the materials from which the Moon coalesced, fractionation of the two elements during core–mantle segregation of the Moon, or crystallization of the lunar magma ocean (LMO). Although the Hf/W ratio of the bulk lunar mantle probably increased slightly as a result of lunar core formation and extraction of an unknown proportion of the siderophile W into the core, recent studies have concluded that the silicate portions of the Earth and Moon had nearly identical Hf/W (see, for example, ref. 19). Consequently, if the Moon formed while ^{182}Hf was still extant, and the silicate portions of the Earth and Moon had identical W isotopic compositions at the time of formation, the isotopic compositions of W would not have evolved to the different compositions observed.

In contrast, fractional crystallization of the LMO almost certainly led to the creation of mantle domains with both higher and lower Hf/W ratios, compared to the bulk lunar mantle. This is due to the more incompatible nature of W in silicate systems, compared with Hf (ref. 20). Crystal–liquid fractionation would, therefore, have led to the creation of ^{182}W-enriched and ^{182}W-depleted domains in the mantle, compared to the ^{182}W of the bulk lunar mantle, if LMO crystallization
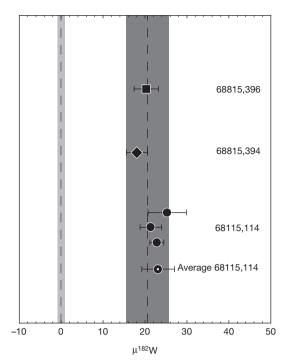
Figure 1 | Values of $\mu^{182}W$ of lunar metals separated from KREEP-rich impact melts analysed by negative thermal ionization mass spectrometry in this study. The data for 68115,114, 68815,394, and 68815,396 are shown as circles, diamond, and square respectively; error bars for our analysis show internal precision of one single measurement, for which the 2 standard deviations (s.d.) external reproducibility is ~4.5 ppm, as demonstrated by replicated standard measurements over the two year period. The white-dotted circle corresponds to the average of the three replicated analyses of 68115,114 metal; error bars show 2 s.d. of these data. The dark grey area and black dashed line indicates the average $\mu^{182}W = +20.6 \pm 5.1$ (2 s.d., n = 3) of the three metal separates from Apollo 16 impact melt rocks analysed here. The light grey dashed line corresponds to the W isotope composition of the modern terrestrial mantle, and the light grey area at $\mu^{182}W = 0$ corresponds to the 2 standard errors (s.e.) uncertainty for repeated analyses of the Alfa Aesar W standard.

was rapid while ^{182}Hf was extant. The comparatively large amount of W needed to make sufficiently high-precision measurements, coupled with sample mass limitations for the Apollo samples, prevented us from making isotopic measurements on rocks derived from lunar mantle domains with different Hf/W from the KREEP source. Two observations, however, suggest that radiogenic ingrowth inside the Moon was not the cause of the ^{182}W-enriched nature of the metals examined here. First, the coupled $^{146,147}Sm$–$^{142,143}Nd$ systematics of crustal rocks derived from the lunar mantle indicate that late stages of LMO crystallization occurred more than 100 Myr after Solar System formation[21], well after ^{182}Hf was extinct. Second, regardless of the timing of LMO crystallization, the mantle source of KREEP was likely to have been a low-Hf/W reservoir, given the W-enriched nature of KREEP. Thus, if the KREEP mantle source rapidly formed during the lifetime of ^{182}Hf, it would have developed a ^{182}W deficit relative to the Earth–Moon system, rather than the observed enrichment (Extended Data Fig. 1), assuming that the mantles of both bodies were in isotopic equilibrium at the time of the Moon's formation. We conclude that

Table 2 | Highly siderophile element contents and Os isotopic compositions of lunar metals

Samples	Re (p.p.b.)	Os (p.p.b.)	Ir (p.p.b.)	Ru (p.p.b.)	Pt (p.p.b.)	Pd (p.p.b.)	$^{187}Os/^{188}Os$	$\pm 2\sigma_{mean}$
68115,113 metal							0.13825	0.00007
Replicate							0.13837	0.00004
68815,394 metal	6.750	57.59	61.99	144.5	167.2	163.6	0.13720	0.00011
68815,396 metal	128.6	1256	1188	2479	3497	2347	0.13480	0.00006

$2\sigma_{mean}$ corresponds to 2 standard errors of an individual $^{187}Os/^{188}Os$ measurement.
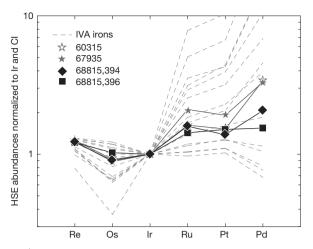
Figure 2 | HSE abundances (normalized to Ir and CI chondrite abundances[33]) of metal separates from sample 68815,394 and 68815,396. The symbols are the same as those used in Fig. 1. Data for IVA irons[15] (grey dashed lines) and impact melts 60315 and 67935 (ref. 16, dark grey stars) are also shown for comparison. (To obtain data for this plot, all obtained element concentrations have been divided by their respective abundance in CI chondrites, then normalized so that they all plot on the same point for Ir.)

the average W isotopic composition of the metals from the Apollo 16 samples is representative of the bulk lunar mantle.

The isotopic similarity between the Earth and Moon for lithophile elements, such as O, Cr, Si and Ti, is well established[22,23], and has led to multiple hypotheses, including the possibility that the giant impact that created the Moon involved an impactor composed of genetically similar materials (that is, similar in terms of nebular components) to the Earth[22], that the Moon was constructed mainly from terrestrial materials rather than the impactor[24], or that the impact and coalescence processes somehow led to thorough isotopic mixing between the two bodies[25]. The recent report of a small offset in the $\Delta^{17}O$ composition of the Moon, compared to the Earth[23], leaves open the possibility that there were small differences in the isotopic compositions of other elements in addition to O. Thus, it is possible that the W isotopic difference between the Earth and the Moon is a result of the latter forming from a mixture of terrestrial and impactor W, assuming that there was no isotopic equilibration following metal segregation. Mixing of W between the Earth's mantle and the impactor mantle, together with varying proportions of impactor metal, could have raised the $\mu^{182}W$ of the Earth's mantle by as much as ~100 p.p.m., if there was negligible equilibration between W from the core of the impactor and the Earth's mantle[26]; or it could have lowered it by as much as ~200 p.p.m., if both the core and mantle of the impactor equilibrated with the Earth's mantle[27]. Thus, even if the mantles of the Earth and the impactor had similar W isotopic compositions at the time of impact, plausible mixtures of terrestrial and impactor W could have produced a Moon with a broad range of possible W isotopic compositions, well beyond the observed offset.

The most parsimonious explanation for the W isotopic offset between the Earth and the Moon is disproportional late accretion to the two bodies. Late accretion is the process whereby substantial mass of materials with chondritic bulk compositions is added to a planetary mantle after core segregation ceases[28]. Late accretion would have led to decreases in the $\mu^{182}W$ values of the mantles of Earth and Moon, because materials with chondritic bulk properties have much higher W concentrations and strongly negative $\mu^{182}W$ values, compared to the terrestrial and lunar mantles. The observed W isotopic offset is consistent with estimates that late accretion added 0.3–0.8 wt% to the mass of the Earth, and ~0.05 wt% to the mass of the Moon[3,4]. The addition of proportionally much greater mass to the Earth, relative to the Moon, has been explained as a consequence of stochastic

processes[29]. If these estimates for late accretion are accurate, then the $\mu^{182}W$ value of the Earth's mantle, before late accretion, was 10–30 p.p.m. higher than at present, compared with only 1–3 p.p.m. higher for the lunar mantle (Extended Data Fig. 2). Because the W isotopic composition of the Moon falls within this very narrow range of isotopic compositions predicted by disproportional late accretion, we conclude that it is the most likely cause of the isotopic offset.

In addition to the requirement that the silicate portions of the Earth and the Moon were the same at the time of the formation of the Moon, an interpretation of disproportional late accretion requires the late accretionary clocks for both the Moon and the Earth to have initiated at the same time. Thus, if significant late accretion to the Earth occurred before the giant impact, metal from the core of the giant impactor must have efficiently stripped the pre-existing HSE from the mantle. It is unknown how much metal–silicate equilibration would be needed to strip the mantle of HSE, or how this level of equilibration would affect W in the mantle. Fluid dynamics experiments and models of metal blobs falling through a magma ocean predict that metallic cores of small impacting bodies efficiently equilibrate with molten silicate, because of fragmentation of the metal cores into droplets small enough to allow rapid metal–silicate mass transfer[30] and turbulent mixing[31]. The nature of merging cores during much larger impact events, particularly with regard to isotopic and elemental equilibration of siderophile elements, is less well understood. It is even possible that elemental metal–silicate equilibration for siderophile elements occurs more rapidly than isotopic equilibration, as has been observed for the HSE Os (ref. 32). Thus, it remains unknown how much the metal from the core of the impactor modified the W isotopic composition of the terrestrial mantle *en route* to the terrestrial core. Knowledge of the W isotopic composition of the terrestrial mantle before the giant impact is needed to estimate accurately the average core–mantle differentiation age of the Earth. Better experimental data on the rates of elemental versus isotopic metal–silicate equilibration will be needed for W in order to tighten existing constraints on the timing of primary Earth differentiation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. Late formation and prolonged differentiation of the Moon inferred from W isotopes in lunar metals. *Nature* **450,** 1206–1209 (2007).
2. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. Tungsten isotopes in ferroan anorthosites: implications for the age of the Moon and the lifetime of its magma ocean. *Icarus* **199,** 245–249 (2009).
3. Day, J. M. D., Pearson, D. G. & Taylor, L. A. Highly siderophile element constraints on accretion and differentiation of the Earth-Moon system. *Science* **315,** 217–219 (2007).
4. Walker, R. J. Highly siderophile elements in the Earth, Moon and Mars: update and implications for planetary accretion and differentiation. *Chem. Erde* **69,** 101–125 (2009).
5. Leya, I., Wieler, R. & Halliday, A. N. Cosmic-ray production of tungsten isotopes in lunar samples and meteorites and its implications for Hf-W cosmochemistry. *Earth Planet. Sci. Lett.* **175,** 1–12 (2000).
6. Kleine, T., Palme, H., Mezger, K. & Halliday, A. N. Hf–W chronometry of lunar metals and the age and early differentiation of the Moon. *Science* **310,** 1671–1674 (2005).
7. Touboul, M. & Walker, R. J. High precision measurement of tungsten isotopes by thermal ionization mass spectrometry. *Int. J. Mass Spectrom.* **309,** 109–117 (2012).
8. Tera, F., Papanastassiou, D. A. & Wasserburg, G. J. Isotopic evidence for a terminal lunar cataclysm. *Earth Planet. Sci. Lett.* **22,** 1–21 (1974).
9. Misra, K. C. & Taylor, L. A. Characteristics of metal particles in Apollo 16 rocks. *Proc. Lunar Sci. Conf.* **6,** 615–639 (1975).
10. Münker, C. *et al.* Evolution of planetary cores and the Earth-Moon system from Nb/Ta systematics. *Science* **301,** 84–87 (2003).
11. Drozd, R. J., Hohenberg, C. M., Morgan, C. J. & Ralston, C. E. Cosmic-ray exposure history at the Apollo 16 and other lunar sites: lunar surface dynamics. *Geochim. Cosmochim. Acta* **38,** 1625–1642 (1974).
12. Lee, D. C., Halliday, A. N., Snyder, G. A. & Taylor, L. A. Age and origin of the Moon. *Science* **278,** 1098–1103 (1997).

13. Norman, M. D., Bennett, V. C. & Ryder, G. Targeting the impactors: siderophile element signatures of lunar impact melts from Serenitatis. *Earth Planet. Sci. Lett.* **202,** 217–228 (2002).

14. Kleine, T., Mezger, K., Münker, C., Palme, H. & Bischoff, A. ^{182}Hf-^{182}W isotope systematics of chondrites, eucrites, and Martian meteorites: chronology of core formation and mantle differentiation in Vesta and Mars. *Geochim. Cosmochim. Acta* **68,** 2935–2946 (2004).

15. McCoy, T. J. *et al.* Group IVA irons: new constraints on the crystallization and cooling history of an asteroidal core with a complex history. *Geochim. Cosmochim. Acta* **75,** 6821–6843 (2011).

16. Fischer-Gödde, M. & Becker, H. Osmium isotope and highly siderophile element constraints on ages and nature of meteoritic components in ancient lunar impact rocks. *Geochim. Cosmochim. Acta* **77,** 135–156 (2012).

17. Kruijer, T. S. *et al.* Protracted core formation and rapid accretion of protoplanets. *Science* **344,** 1150–1154 (2014).

18. Sharp, M. *et al.* Characterization of the dominant impactor signature for Apollo 17 impact melt rocks. *Geochim. Cosmochim. Acta* **131,** 62–80 (2014).

19. König, S. *et al.* The Earth's tungsten budget during mantle melting and crust formation. *Geochim. Cosmochim. Acta* **75,** 2119–2136 (2011).

20. Righter, K. & Shearer, C. K. Magmatic fractionation of Hf and W: constraints on the timing of core formation and differentiation in the Moon and Mars. *Geochim. Cosmochim. Acta* **67,** 2497–2507 (2003).

21. McLeod, C. L., Brandon, A. D. & Armytage, R. M. G. Constraints on the formation age and evolution of the Moon from ^{142}Nd-^{143}Nd systematics of Apollo 12 basalts. *Earth Planet. Sci. Lett.* **396,** 179–189 (2014).

22. Dauphas, N., Burkhardt, C., Warren, P. H. & Teng, F.-Z. Geochemical arguments for an Earth-like Moon-forming impactor. *Phil. Trans. R. Soc. A* **372,** 20130244 (2014).

23. Herwartz, D., Pack, A., Friedrichs, B. & Bischoff, A. Identification of the giant impactor Theia in lunar rocks. *Science* **344,** 1146–1150 (2014).

24. Ćuk, M. & Stewart, S. T. Making the Moon from a fast-spinning Earth: a giant impact followed by resonant despinning. *Science* **338,** 1047–1052 (2012).

25. Canup, R. M. Forming a Moon with an Earth-like composition via a giant impact. *Science* **338,** 1052–1055 (2012).

26. Halliday, A. N. Mixing, volatile loss and compositional change during impact-driven accretion of the Earth. *Nature* **427,** 505–509 (2004).

27. Halliday, A. N. A young Moon-forming giant impact at 70–110 million years accompanied by late-stage mixing, core formation and degassing of the Earth. *Phil. Trans. R. Soc. A* **366,** 4163–4181 (2008).

28. Chou, C.-L. Fractionation of siderophile elements in the Earth's upper mantle. *Proc. Lunar Planet. Sci. Conf.* **9,** 219–230 (1978).

29. Bottke, W. F., Walker, R. J., Day, J. M. D., Nesvorny, D. & Elkins-Tanton, L. Stochastic late accretion to Earth, the Moon, and Mars. *Science* **330,** 1527–1530 (2010).

30. Stevenson, D. J. in *Origin of the Earth* (eds Newsom, H. E. & Drake, J. H.) 231–249 (Oxford Univ. Press, 1990).

31. Deguen, R., Landeau, M. & Olson, P. Turbulent metal–silicate mixing, fragmentation, and equilibration in magma oceans. *Earth Planet. Sci. Lett.* **391,** 274–287 (2014).

32. Yokoyama, T., Walker, D. & Walker, R. J. Low osmium solubility in silicate at high pressures and temperatures. *Earth Planet. Sci. Lett.* **279,** 165–173 (2009).

33. Horan, M. F., Walker, R. J., Morgan, J. W., Grossman, J. N. & Rubin, A. E. Highly siderophile elements in chondrites. *Chem. Geol.* **196,** 27–42 (2003).

Author Contributions M.T. conducted the W isotopic measurements and was involved in both the interpretations and the writing of the manuscript. I.S.P. conducted the measurements of the highly siderophile elements and Os isotopes, and was involved in both the interpretations and the writing of the manuscript. R.J.W. was involved in both the interpretations and the writing of the manuscript.

Author Information The data presented here can be found in the EarthChem library (http://www.earthchem.org/library/browse/view?id=849). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.T. (mathieu.touboul@ens-lyon.fr), I.S.P. (ipuchtel@umd.edu) or R.J.W. (rjwalker@umd.edu).

METHODS

Samples 68115 and 68815 are glassy polymict breccias composed of a variety of impact melts, as well as relict aluminous plagioclase clasts. A 3.2 g sample split (68115,114), a 2.7 g sample split (68815,394) and a 3 g sample split (68815,396) appeared visually free of large clasts, and were selected for study. Each sample was crushed in an agate mortar and separated into several size fractions using nylon sieves. Magnetic fractions were separated using a hand-magnet, and further purified by repeated grinding, magnetic separation and ultrasonication in high-purity water. About 100 mg, ~40 mg and ~20 mg of separated, high-purity metal, were obtained for 68115,114, 68815,394 and 68815,396, respectively.

After metal dissolution in 6 M HCl, a ~0.5% aliquot of each sample was spiked for determination of Hf and W concentrations. Spike-sample mixtures were equilibrated in 7 ml screw-cap Teflon vials at 130 °C for 2 days, then the solutions were evaporated to dryness. Residues were then re-dissolved in 2 ml of an HCl (0.5 M)–HF (0.5 M) mixture, and W and Hf were then purified using a previously established anion exchange chromatography technique[1,2].

Tungsten was separated from the remaining sample solution (99.5%) using anion exchange chromatography for determination of isotopic composition. After evaporation, residues were digested twice in concentrated HNO_3, with traces of H_2O_2, over ~24 h at 120 °C, and evaporated to dryness. Residues were then converted into the chloride form by repeated dissolution in 6 M HCl and subsequent evaporation. The samples were finally dissolved in 2 ml of an HCl (0.5 M)–HF (0.5 M) mixture and then purified using the three-step anion exchange chromatography described in ref. 7. Approximately 2.5, 0.7 and 0.5 μg of W were ultimately harvested from the metal fractions of 68115,114, 68815,394 and 68815,396, respectively, with corresponding procedural yields of ~85%. These quantities of W allowed us to make three independent high-precision W isotope measurements for 68115,114, one for 68815,394 and one for 68815,396. The total procedural W blank was 1 ± 0.5 ng, and was negligible. Tungsten isotope compositions were measured to <5 p.p.m. precision level using a Thermo Triton thermal ionization mass spectrometer at the University of Maryland, following our published analytical procedure[7]. A similar level of reproducibility for W separated from cosmochemical metals is demonstrated through duplicate and triplicate analyses of similar quantities of W extracted from iron meteorites as reported in supplementary table 5 in ref. 17.

For determining Os isotopic compositions and HSE concentrations, 1.1 mg and 0.61 mg of metal in the form of HCl solution from 68115,114 and 68815,396, respectively, and 1.1 mg of a mixture of silicate and metal separate from 68815,394, together with 5 ml of triple-distilled, concentrated HNO_3, 4 ml of triple-distilled, concentrated HCl, and appropriate amounts of mixed ^{185}Re–^{190}Os and HSE (^{99}Ru, ^{105}Pd, ^{191}Ir, ^{194}Pt) spikes, were sealed in double-cleaned, chilled 25 ml Pyrex borosilicate Carius tubes and heated to 270 °C for at least 96 h. Osmium was extracted from the acid solution by CCl_4 solvent extraction[34], then back-extracted into HBr, followed by purification using microdistillation[35]. Iridium, Ru, Pt, Pd and Re were separated and purified using anion exchange chromatography.

The total analytical blanks in picograms were as follows: Re, 0.16; Os, 0.41; Ir, 0.31; Ru, 4.5; Pt, 95; Pd, 5.3. The abundances of the highly siderophile elements, and the $^{187}Os/^{188}Os$ ratios, were corrected using the values for the blank measured with this set of samples.

Osmium isotopic measurements were accomplished by negative thermal ionization mass spectrometry (NTIMS)[36]. All samples were analysed using a secondary electron multiplier detector of a Thermo Fisher Triton mass spectrometer at the Isotope Geochemistry Laboratory (IGL), University of Maryland. The measured isotopic ratios were corrected for mass fractionation using $^{192}Os/^{188}Os = 3.083$. The internal precision of measured $^{187}Os/^{188}Os$ for all samples was better than 0.1% relative. The $^{187}Os/^{188}Os$ of 300–500 pg loads of the in-house Johnson-Matthey Os standard measured during the period of the analytical campaign averaged 0.11376 ± 10 ($2\sigma_{stdev}$, $N = 64$). This value characterizes the external precision of the isotopic analysis (0.1%), which we use to calculate the true uncertainty on the measured $^{187}Os/^{188}Os$ ratio for each individual sample. The measured $^{187}Os/^{188}Os$ ratios further were also corrected for the instrumental bias relative to the average $^{187}Os/^{188}Os = 0.11379$ measured for the Johnson-Matthey Os standard on the Faraday cups of the IGL Triton. The correction factor of 1.00026 was calculated by dividing this value by the average $^{187}Os/^{188}Os$ measured in the Johnson-Matthey Os standard on the secondary electron multiplier of the same instrument.

The measurements of Ru, Pd, Re, Ir and Pt were performed at the IGL by inductively coupled plasma mass-spectrometry (ICP-MS) using a Nu Plasma instrument with a triple electron multiplier configuration in a static mode. Isotopic mass fractionation was monitored and corrected for by interspersing samples and standards. The accuracy of the data was assessed by comparing the results for the reference materials UB-N and GP-13 obtained during the ongoing analytical campaign[37] with the results from other laboratories. Concentrations of all HSE and Os isotopic compositions obtained at the IGL are in good agreement with the other laboratories. Diluted spiked aliquots of iron meteorites were run during each analytical session as secondary standards. The results from these runs agreed within 0.5% for Re and Ir, and within 2% for Ru, Pt and Pd, with fractionation-corrected values obtained from measurements of undiluted iron meteorites using Faraday cups of the same instrument with a signal of >100 mV for the minor isotopes. Depending on the total amount of HSE aliquant processed for each sample, the uncertainties on the HSE concentrations varied between 0.5% and 2% for Re, 0.1% and 0.6% for Os, 4% and 50% for Pt, 2% and 3% for Pd, and were 0.5% for Ir, and 2% for Ru.

Sample size. We had prior data for W concentrations in similar metals, and these data were used to define sample size.

34. Cohen, A. S. & Waters, F. G. Separation of osmium from geological materials by solvent extraction for analysis by thermal ionisation mass spectrometry. *Anal. Chim. Acta* **332,** 269–275 (1996).
35. Birck, J. L., Roy-Barman, M. & Capman, F. Re-Os isotopic measurements at the femtomole level in natural samples. *Geostand. Newsl.* **21,** 19–27 (1997).
36. Creaser, R. A., Papanastassiou, D. A. & Wasserburg, G. J. Negative thermal ion spectrometry of osmium, rhenium, and iridium. *Geochim. Cosmochim. Acta* **55,** 397–401 (1991).
37. Puchtel, I. S., Walker, R. J., Touboul, M., Nisbet, E. G. & Byerly, G. R. Insights into early Earth from the Pt-Re-Os isotope and highly siderophile element abundance systematics of Barberton komatiites. *Geochim. Cosmochim. Acta* **125,** 394–413 (2014).
38. Arevalo, R. & McDonough, W. F. Tungsten geochemistry and implications for understanding the Earth's interior. *Earth Planet. Sci. Lett.* **272,** 656–665 (2008).

LETTER

Lunar tungsten isotopic evidence for the late veneer

Thomas S. Kruijer[1], Thorsten Kleine[1], Mario Fischer-Gödde[1] & Peter Sprung[1]†

According to the most widely accepted theory of lunar origin, a giant impact on the Earth led to the formation of the Moon, and also initiated the final stage of the formation of the Earth's core[1]. Core formation should have removed the highly siderophile elements (HSE) from Earth's primitive mantle (that is, the bulk silicate Earth), yet HSE abundances are higher than expected[2]. One explanation for this overabundance is that a 'late veneer' of primitive material was added to the bulk silicate Earth after the core formed[2]. To test this hypothesis, tungsten isotopes are useful for two reasons: first, because the late veneer material had a different $^{182}W/^{184}W$ ratio to that of the bulk silicate Earth, and second, proportionally more material was added to the Earth than to the Moon[3]. Thus, if a late veneer did occur, the bulk silicate Earth and the Moon must have different $^{182}W/^{184}W$ ratios. Moreover, the Moon-forming impact would also have created ^{182}W differences because the mantle and core material of the impactor with distinct $^{182}W/^{184}W$ would have mixed with the proto-Earth during the giant impact. However the $^{182}W/^{184}W$ of the Moon has not been determined precisely enough to identify signatures of a late veneer or the giant impact. Here, using more-precise measurement techniques, we show that the Moon exhibits a ^{182}W excess of 27 ± 4 parts per million over the present-day bulk silicate Earth. This excess is consistent with the expected ^{182}W difference resulting from a late veneer with a total mass and composition inferred from HSE systematics[2]. Thus, our data independently show that HSE abundances in the bulk silicate Earth were established after the giant impact and core formation, as predicted by the late veneer hypothesis. But, unexpectedly, we find that before the late veneer, no ^{182}W anomaly existed between the bulk silicate Earth and the Moon, even though one should have arisen through the giant impact. The origin of the homogeneous ^{182}W of the pre-late-veneer bulk silicate Earth and the Moon is enigmatic and constitutes a challenge to current models of lunar origin.

The extinct ^{182}Hf–^{182}W system (half-life, 8.9 Myr) is a versatile tool for investigating potential isotopic differences between the Earth and Moon because it is sensitive to different degrees of metal–silicate equilibration during core formation, to mixing processes during the giant impact, and to the addition of meteoritic material to silicate mantles after the giant impact. During core formation, the lithophile Hf is fractionated from the siderophile W, and the resulting distinct Hf/W ratio causes the mantle and the core to have different values of $\varepsilon^{182}W$ (that is, the deviation in parts per 10,000 of the $^{182}W/^{184}W$ ratio from the value of the present-day bulk silicate Earth), depending on the timescales and conditions of core formation[4]. This makes $\varepsilon^{182}W$ a sensitive tracer of proto-Earth and impactor components in the Moon, as well as of the late accretion of chondritic, ^{182}W-depleted material that was added to the mantles of the Earth and Moon. For instance, small ^{182}W excesses in some Archaean terrestrial rocks may reflect mantle sources that lack a portion of the late veneer[5], although such ^{182}W heterogeneities may also result from early mantle differentiation processes[6]. As the fraction of late-accreted mass added to the lunar mantle was much smaller than the fraction added to the bulk silicate Earth (BSE)[3,7], mass balance calculations suggest that the late

veneer would inevitably have generated an Earth–Moon ^{182}W difference[8]. A $\varepsilon^{182}W$ difference between the Moon and the BSE may also exist because, during the giant impact that gave rise to the Moon, different proportions of the impactor material, which had distinct $\varepsilon^{182}W$, ended up in the Earth and the Moon. Therefore, identifying a difference of $\varepsilon^{182}W$ (or lack thereof) between the Earth and Moon would provide key information regarding the earliest history of the Earth–Moon system; such information would constrain the process and timing of late accretion and help shape models of lunar origin.

However, determining the lunar $\varepsilon^{182}W$ is complicated by cosmic-ray-induced secondary neutron capture reactions. These reactions not only involve ^{182}W production via neutron capture by ^{181}Ta, but also neutron-capture-induced burnout of ^{182}W (ref. 9). Hence, the previously measured value of $\varepsilon^{182}W = 0.09 \pm 0.10$ (2 standard error, s.e.) for the Moon[10], which was based on Ta-free and thus cosmogenic-^{182}W-free lunar metals, may have been lowered by ^{182}W burnout, and therefore only provides a minimum estimate. Without a suitable neutron dosimeter, the previously measured $\varepsilon^{182}W$ values of the lunar metals had been corrected using cosmic ray exposure ages[10,11], but these do not directly measure the neutron dose that affects W isotopes. Furthermore, the analytical precision of the previous studies was >10 p.p.m. (2 s.e.) for individual samples, which is insufficient to resolve a (hypothetical) small ^{182}W anomaly of the Moon. Consequently, it has been unclear until now whether the Moon and the present-day BSE differ significantly in $\varepsilon^{182}W$.

In this study, we determined the $\varepsilon^{182}W$ of the Moon using improved analytical techniques for high-precision W isotope measurements[12] combined with a new approach to quantifying cosmogenic $\varepsilon^{182}W$ variations using Hf isotopes[13]. All investigated samples are impact rocks (see ref. 14 and references therein), so their $\varepsilon^{182}W$ may have been modified through the addition of meteoritic material. We determined the magnitude of such meteorite contamination from the abundances of HSE in the investigated samples[15,16], producing corrections of ~0.04 $\varepsilon^{182}W$ for samples 68115 and 68815, and of less than ~0.02 $\varepsilon^{182}W$ for all other samples (Table 1). We focused on KREEP-rich samples because they have a near-constant Ta/W ratio (ref. 11), implying essentially invariant effects on $\varepsilon^{182}W$ for a given neutron dose (KREEP is thought to represent the residual liquid of the lunar magma ocean and is enriched in incompatible elements including potassium, K, rare-earth elements, REE, and phosphorus, P). The KREEP-rich samples exhibit a well-defined $\varepsilon^{182}W$–$\varepsilon^{180}Hf$ correlation (Fig. 1), reflecting the fact that Ta, W and Hf isotopes are most sensitive to neutrons of similar (epithermal) energies. All investigated KREEP-rich samples have a common pre-exposure $\varepsilon^{182}W$ (that is, unaffected by neutron capture) of +0.27 ± 0.04 (95% confidence interval) defined either by the intercept of the $\varepsilon^{182}W$–$\varepsilon^{180}Hf$ correlation (Fig. 1) or by samples lacking significant $\varepsilon^{180}Hf$ anomalies (samples 14321, 68115, 68815; Fig. 2). We interpret the pre-exposure $\varepsilon^{182}W$ of the KREEP-rich samples to represent that of the bulk silicate Moon, because lunar differentiation at ~4.4 Gyr ago (Ga; refs 13, 17, 18) was too late to produce ^{182}W variations within the Moon, consistent with the indistinguishable $\varepsilon^{182}W$ of non-irradiated mare basalts[13] and KREEP (see Methods). Our newly determined pre-exposure $\varepsilon^{182}W$

[1]Institut für Planetologie, Westfälische Wilhelms-Universität Münster, Wilhelm-Klemm-Strasse 10, D-48149 Münster, Germany. †Present address: Universität zu Köln, Institut für Geologie und Mineralogie, Zülpicher Strasse 49b, D-50674 Köln, Germany.

Table 1 | Tungsten and Hf isotope data for KREEP-rich samples analysed by MC-ICPMS

Sample	t_{CRE} § (Ma)	N	$\varepsilon^{182/184}$W (6/4)$_{meas.}$* ($\pm2\sigma$)	$\varepsilon^{182/184}$W (6/4)$_{corr.}$† ($\pm2\sigma$)	W$_{met.}$‡ (%)^c	ε^{180}Hf ($\pm$95% confidence interval)
Weakly irradiated samples						
14321, 1827		2	0.29±0.10	0.29±0.10	0.1	−0.02±0.08
14321, 1856		6	0.27±0.05	0.27±0.05	0.1	−0.02±0.08
14321 (weighted mean)	23.8		0.27±0.04	0.27±0.04	0.1	
68115, 295		4	0.24±0.06	0.28±0.06	1.4	0.02±0.13
68115, 112		3	0.27±0.10	0.31±0.10	1.4	0.02±0.13
68115 (weighted mean)	2.08		0.25±0.05	0.29±0.05	1.4	
68815, 400	2.04	2	0.18±0.10	0.21±0.10	1.0	−0.03±0.16
14321, 68115, 68815 (weighted mean)			0.25±0.03	0.27±0.03		
Strongly irradiated samples						
14163, 921	NA	5	2.35±0.04	ND	ND	−3.79±0.08
12034, 120	NA	3	1.26±0.10	ND	ND	−1.79±0.09
14310, 676	259	3	1.87±0.10	ND	ND	−2.94±0.08
62235, 122	153	2	1.63±0.10	ND	ND	−2.46±0.10

All Hf isotope data are from ref. 13, except for samples 14321 and 14163 which were newly analysed (see Methods). N, number of measurements of each sample; NA, not available; ND, not determined.
* Measured ε^{182}W internally normalized to ^{186}W/^{184}W = 0.92767, denoted by (6/4).
† ε^{182}W corrected for meteoritic contamination on the lunar surface using measured HSE and W abundances (see Methods).
‡ Percentage of W in sample that derives from meteoritic impactor component added at lunar surface.
§ Cosmic-ray exposure ages (t_{CRE}) of lunar samples (ref. 14 and references therein).

of $+0.27 \pm 0.04$ is significantly higher than the previously obtained mean value of 0.09 ± 0.10 for lunar metal samples (ref. 10), but for non-irradiated samples (68115, 68815) there is good agreement between our data and previous data (Fig. 2). For more strongly irradiated samples, however, the ε^{182}W of the metals tends to be slightly lower[10], resulting in an overall decrease of the mean ε^{182}W inferred from the lunar metals. Therefore, the higher pre-exposure ε^{182}W of $+0.27 \pm 0.04$ determined here reflects not only the better precision of our measurements, but also that the previous study[10] did not fully quantify neutron capture effects in the metals.

The well-resolved ^{182}W excess of the Moon compared to the present-day BSE (Fig. 2) places important constraints on the occurrence, mass and timing of the late veneer as well as on the origin of the Moon. Below we first evaluate the magnitude of any ε^{182}W difference between the BSE and the Moon induced by the late veneer, and then we assess whether there is a resolvable ^{182}W anomaly in the Moon resulting from the mixing of impactor and proto-Earth material during the giant impact. The mass and composition of the late veneer is constrained through absolute and relative HSE abundances and ratios of S, Se and Te in Earth's primitive mantle[2,19,20]. On this basis, the late veneer probably had a carbonaceous-chondrite-like composition with a minor fraction of iron-meteorite-like material[16], corresponding to

~0.35% of Earth's mass. This composition can explain several geochemical signatures of the Earth's mantle, including its chondritic Os/Ir, Pt/Ir and Rh/Ir but suprachondritic Ru/Ir and Pd/Ir, as well as its ^{187}Os/^{188}Os value[2] and Se–Te systematics[19]. Mass balance considera-
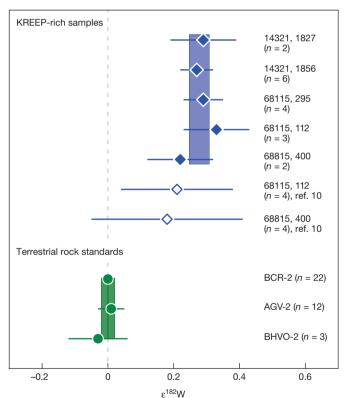
Figure 2 | ε^{182}W data of KREEP-rich samples and terrestrial rock standards. Top panel, data from this study (filled symbols) and for metal samples from ref. 10 (open symbols). Data points of 68115 and 68815 (this study) were corrected for a minor contribution from meteoritic contamination at the lunar surface (Table 1). Error bars indicate external uncertainties derived from the 2 s.d. obtained for terrestrial rock standards analysed in this study (if $N < 4$) or 95% confidence interval of multiple solution replicates of a sample (if $N \geq 4$) (Extended Data Table 1). Bottom panel, data from terrestrial rock standards. Top panel, weighted mean ($n = 5$) ε^{182}W = $+0.27 \pm 0.03$ (95% confidence interval, blue shaded area); bottom panel, mean ε^{182}W ($n = 37$) = 0.00 ± 0.10 (2 s.d.) = 0.00 ± 0.02 (95% confidence interval, green shaded area). ε^{182}W has been internally normalized to ^{186}W/^{184}W = 0.92767, and dashed grey line shows ε^{182}W = 0.

Figure 1 | Plot of ε^{182}W versus ε^{180}Hf determined for KREEP-rich samples. ε^{182}W has been internally normalized to ^{186}W/^{184}W = 0.92767: elsewhere this is referred to as ε^{182}W (6/4) (see Methods and Table 1). Solid line is a best-fit linear regression through the data (slope = -0.549 ± 0.019; MSWD = 0.36) with the intersection at ε^{180}Hf = 0 (arrowed) defining the pre-exposure ε^{182}W (= $+0.27 \pm 0.04$, $\pm$95% confidence interval). Error bars, external uncertainties (95% confidence interval or 2 s.d.; Extended Data Table 2).

Figure 3 | Plot of $\varepsilon^{182}W$ versus mass fraction of late-accreted material on Earth. Mass balance calculations (blue envelope) predict a positive $\varepsilon^{182}W$ signature if the full complement of late accretion is subtracted from the present-day BSE composition (right red circle, $\varepsilon^{182}W \equiv 0$), in excellent agreement with the lunar $\varepsilon^{182}W$ value (left red circle; $\varepsilon^{182}W = +0.27 \pm 0.04$, 95% confidence interval). The uncertainty on the pre-late veneer $\varepsilon^{182}W$ of the BSE mainly results from the uncertainty on the W concentration of the BSE (13 ± 5 p.p.b.). For details about the late veneer composition used in the mass balance, see Methods.

tions imply that the addition of a late veneer of this composition lowered the $\varepsilon^{182}W$ of the BSE by $+0.22^{+0.16}_{-0.06}$ (Fig. 3), where this estimate's uncertainty mainly comes from that of the W concentration of the BSE. A late veneer composed exclusively of known groups of chondrites would have resulted in an only slightly smaller but still consistent shift of ~0.1 to ~0.3 $\varepsilon^{182}W$ (see Methods). The corresponding effect of the late veneer on the lunar $\varepsilon^{182}W$ is negligible, given that the mass fraction added to the lunar mantle was an order of magnitude smaller than that added to the Earth[3,7]. Therefore, the $\varepsilon^{182}W$ difference between the Moon and the present-day BSE of $+0.27 \pm 0.04$ can be entirely accounted for by the addition of ^{182}W-depleted material to the BSE during late accretion, with a total mass consistent with that derived from the HSE abundances in Earth's mantle[2,16] (Fig. 3). This implies that previously accumulated HSEs in the Earth's mantle had been sequestered into Earth's core during the giant impact[8], demonstrating that the entire late veneer was added after the giant impact and the final stages of core segregation. These ^{182}W results therefore provide independent evidence for the late veneer hypothesis by demonstrating that the HSE abundances in the Earth's mantle were established by addition of primitive material after the Earth's core formed.

The close agreement between the predicted late-veneer-induced $\varepsilon^{182}W$ shift and that observed between the BSE and the Moon suggests that, before addition of the late veneer, the BSE and the Moon had indistinguishable values of $\varepsilon^{182}W$. This implies that there is no resolvable radiogenic ^{182}W difference between the Moon and the Earth, probably because the Moon formed late[10,18,21]—when ^{182}Hf was already extinct—or perhaps because the BSE and the silicate Moon have very similar Hf/W. In addition, the data suggest that the giant impact did not induce a resolvable ^{182}W anomaly in the Moon. This is consistent with the Earth–Moon isotopic homogeneity observed for Ti (ref. 22), Si (ref. 23) and O (ref. 24). We note that a small O isotope difference exists between the Moon and the present-day BSE and this is most probably caused by late accretion, because the late veneer was probably dominated by carbonaceous-chondrite-like material (see above) and, consequently, had a distinctly lower $\Delta^{17}O$ than the Earth and Moon. Thus, addition of a late veneer with this composition and a mass as derived from HSE and ^{182}W systematics would have led to the observed Earth–Moon difference in $\Delta^{17}O$ (ref. 24). So as is the case for $\varepsilon^{182}W$, once the effect of late accretion is taken into account, the pre-late-veneer BSE and the Moon have indistinguishable $\Delta^{17}O$.

The Earth–Moon isotopic homogeneity for Ti, Si and O may mean that the Earth and the impactor accreted from a homogeneous inner-disk reservoir[25] or that the Moon formed either from proto-Earth mantle material[26,27], or from equal portions of the mantles of two colliding half-Earths[28]. However, in both cases, the homogeneous $\varepsilon^{182}W$ of the pre-late-veneer BSE and the Moon is not easily explained, because the giant impact would have modified the $\varepsilon^{182}W$ of proto-Earth's mantle in two ways: (1) by adding the impactor mantle material, which presumably had distinct $\varepsilon^{182}W$, and (2) through the (partial) equilibration of the impactor core with the proto-Earth's mantle. As such, the pre-giant-impact $\varepsilon^{182}W$ of the Earth's mantle was most probably different from its post-giant-impact value, meaning that making the Moon out of the proto-Earth's mantle would probably not result in a homogeneous $\varepsilon^{182}W$ of the pre-late-veneer BSE and the Moon (Extended Data Fig. 3). Moreover, the lunar accretion disk probably would have contained W-rich but ^{182}W-depleted impactor core material[1,26,28], which consequently would have generated a significant shift in the W isotope composition of the proto-lunar material (Extended Data Fig. 4). Thus, although specific impactor compositions, proto-Earth compositions, and impact conditions that could produce a similar ^{182}W composition of the pre-late-veneer BSE and the Moon can be identified[25], it is far more likely that significant Earth–Moon $\varepsilon^{182}W$ differences were produced, even if the Moon predominantly consists of proto-Earth material. Consequently, it would take extraordinary circumstances to generate the Earth–Moon $\varepsilon^{182}W$ homogeneity through simple mixing of proto-Earth and impactor material during the giant impact.

The $\varepsilon^{182}W$ homogeneity might be accounted for by post-giant-impact equilibration via a shared atmosphere of the lunar accretion disk and the Earth's mantle[29], but this would be difficult for a refractory element like W (ref. 22). Other possible origins of the $\varepsilon^{182}W$ homogeneity might be that the Moon was formed through impact-triggered fission from a fast-spinning proto-Earth[26] or that efficient equilibration occurred during the collision of two half-Earths[28]. However, in the last two scenarios it still has to be evaluated whether equilibration between the impactor and the proto-Earth would be possible before ejection of the proto-lunar material. Either way, the Earth–Moon ^{182}W homogeneity constitutes a fundamental constraint on any successful model of lunar origin.

1. Canup, R. M. & Asphaug, E. Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature* **412**, 708–712 (2001).
2. Walker, R. J. Highly siderophile elements in the Earth, Moon and Mars: update and implications for planetary accretion and differentiation. *Chem. Erde Geochem.* **69**, 101–125 (2009).
3. Day, J. M. D., Pearson, D. G. & Taylor, L. A. Highly siderophile element constraints on accretion and differentiation of the Earth-Moon system. *Science* **315**, 217–219 (2007).
4. Kleine, T. *et al.* Hf–W chronology of the accretion and early evolution of asteroids and terrestrial planets. *Geochim. Cosmochim. Acta* **73**, 5150–5188 (2009).
5. Willbold, M., Elliott, T. & Moorbath, S. The tungsten isotopic composition of the Earth's mantle before the terminal bombardment. *Nature* **477**, 195–198 (2011).
6. Touboul, M., Puchtel, I. S. & Walker, R. J. ^{182}W evidence for long-term preservation of early mantle differentiation products. *Science* **335**, 1065–1069 (2012).
7. Day, J. M. D., Walker, R. J., James, O. B. & Puchtel, I. S. Osmium isotope and highly siderophile element systematics of the lunar crust. *Earth Planet. Sci. Lett.* **289**, 595–605 (2010).
8. Walker, R. J. Siderophile element constraints on the origin of the Moon. *Phil. Trans. R. Soc. A* **372**, http://dx.doi.org/10.1098/rsta.2013.0258 (2014).
9. Leya, I., Wieler, R. & Halliday, A. N. Cosmic-ray production of tungsten isotopes in lunar samples and meteorites and its implications for Hf–W cosmochemistry. *Earth Planet. Sci. Lett.* **175**, 1–12 (2000).
10. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. Late formation and prolonged differentiation of the Moon inferred from W isotopes in lunar metals. *Nature* **450**, 1206–1209 (2007).
11. Kleine, T., Palme, H., Mezger, K. & Halliday, A. N. Hf-W chronometry of lunar metals and the age and early differentiation of the Moon. *Science* **310**, 1671–1674 (2005).

12. Kruijer, T. S. *et al.* Protracted core formation and rapid accretion of protoplanets. *Science* **344,** 1150–1154 (2014).
13. Sprung, P., Kleine, T. & Scherer, E. E. Isotopic evidence for chondritic Lu/Hf and Sm/Nd of the Moon. *Earth Planet. Sci. Lett.* **380,** 77–87 (2013).
14. Meyer, C. The Lunar Sample Compendium. Available at http://curator.jsc.nasa. gov/Lunar/lsc/index.cfm (20 May 2013).
15. Morgan, J. W., Ganapathy, R., Higuchi, H., Krähenbühl, U. & Anders, E. Lunar basins: tentative characterization of projectiles, from meteoritic elements in Apollo 17 boulders. *Proc. Lunar Planet. Sci. Conf.* **2,** 1703–1736 (1974).
16. Fischer-Gödde, M. & Becker, H. Osmium isotope and highly siderophile element constraints on ages and nature of meteoritic components in ancient lunar impact rocks. *Geochim. Cosmochim. Acta* **77,** 135–156 (2012).
17. Gaffney, A. M. & Borg, L. E. A young solidification age for the lunar magma ocean. *Geochim. Cosmochim. Acta* **140,** 227–240 (2014).
18. Carlson, R. W., Borg, L. E., Gaffney, A. M. & Boyet, M. Rb-Sr, Sm-Nd and Lu-Hf isotope systematics of the lunar Mg-suite: the age of the lunar crust and its relation to the time of Moon formation. *Phil. Trans. R. Soc. A* 372, http://dx.doi.org/ 10.1098/rsta.2013.024 (2014).
19. Wang, Z. & Becker, H. Ratios of S, Se and Te in the silicate Earth require a volatile-rich late veneer. *Nature* **499,** 328–331 (2013).
20. Becker, H. *et al.* Highly siderophile element composition of the Earth's primitive upper mantle: constraints from new data on peridotite massifs and xenoliths. *Geochim. Cosmochim. Acta* **70,** 4528–4550 (2006).
21. Borg, L. E., Connelly, J. N., Boyet, M. & Carlson, R. W. Chronological evidence that the Moon is either young or did not have a global magma ocean. *Nature* **477,** 70–72 (2011).
22. Zhang, J., Dauphas, N., Davis, A. M., Leya, I. & Fedkin, A. The proto-Earth as a significant source of lunar material. *Nature Geosci.* **5,** 251–255 (2012).
23. Armytage, R. M. G., Georg, R. B., Williams, H. M. & Halliday, A. N. Silicon isotopes in lunar rocks: implications for the Moon's formation and the early history of the Earth. *Geochim. Cosmochim. Acta* **77,** 504–514 (2012).
24. Herwartz, D., Pack, A., Friedrichs, B. & Bischoff, A. Identification of the giant impactor Theia in lunar rocks. *Science* **344,** 1146–1150 (2014).
25. Dauphas, N., Burkhardt, C., Warren, P. & Teng, F.-Z. Geochemical arguments for an Earth-like Moon-forming impactor. *Phil. Trans. R. Soc. A* http://dx.doi.org/ 10.1098/rsta.2013.0244 (2014).
26. Ćuk, M. & Stewart, S. T. Making the Moon from a fast-spinning Earth: a giant impact followed by resonant despinning. *Science* **338,** 1047–1052 (2012).
27. Reufer, A., Meier, M. M. M., Benz, W. & Wieler, R. A hit-and-run giant impact scenario. *Icarus* **221,** 296–299 (2012).
28. Canup, R. M. Forming a Moon with an Earth-like composition via a giant impact. *Science* **338,** 1052–1055 (2012).
29. Pahlevan, K. & Stevenson, D. J. Equilibration in the aftermath of the lunar-forming giant impact. *Earth Planet. Sci. Lett.* **262,** 438–449 (2007).

METHODS

Sample preparation and chemical separation of W. Seven KREEP-rich whole rock samples that span a wide range in cosmic ray exposure ages were selected for this study (12034, 14163, 14310, 14321, 62235, 68115, 68815)[14]. The Hf isotope compositions of these samples had been determined by us previously[13]. Only the Hf isotope compositions of samples 14321 and 14163 were newly measured (Table 1) by the same methods used in the previous study[13]. Note that all lunar ε^{180}Hf and ε^{178}Hf data are given relative to the average composition obtained for terrestrial rock standards. For W isotope analyses, samples received as rock fragments were ultrasonically cleaned and rinsed with ethanol, then carefully crushed and ground to a fine powder in an agate mortar. Combined W and Hf isotope compositions were determined on separate splits of the same powder aliquots ($\sim$0.2 g). From three samples with no resolvable Hf isotope anomaly (68115, 68815, 14321) more material ($\sim$0.9 g) was powdered for additional W isotope measurements. The analytical techniques for sample digestion, chemical separation of W, and W isotope ratio measurements by MC-ICPMS (Multi Collector Inductively Coupled Plasma Mass Spectrometry) are based on our previously developed procedures[12,30]. The KREEP-rich samples and terrestrial rock standards were dissolved by table-top digestion in $\sim$20 ml HF–HNO$_3$ (2:1) at 130–150 °C for 2–3 days. After digestion, samples were evaporated to dryness at 130 °C, and then concentrated HNO$_3$-HCl was added repeatedly to the samples to remove fluoride precipitates. Subsequently, the samples were converted through repeated evaporations to dryness in 6 M HCl–0.06 M HF. Finally, the samples were completely dissolved in 25 ml 6 M HCl–0.06 M HF. For three weakly irradiated samples (BL01-03) approximately 1–10% aliquots (equivalent to $\sim$2–4 ng W) of these 6 M HCl–0.06 M HF solutions were spiked with a mixed ^{180}Hf–^{183}W tracer that was calibrated against pure Hf and W metal standards[31]. Tungsten was separated from the sample matrix using a two-stage anion exchange chromatography technique that was slightly modified from previously published methods[30–33]. All spike-free fractions were evaporated to dryness and re-dissolved in 25 ml 0.5 M HCl–0.5 M HF before loading onto the first anion exchange column (4 ml BioRad AG1 × 8, 200–400 mesh). The sample matrix was rinsed off the column in 10 ml 0.5 M HCl–0.5 M HF, followed by another rinse with 10 ml 8 M HCl–0.01 M HF in which significant amounts of Ti, Zr and Hf were eluted. Finally, W was eluted in 15 ml 6 M HCl–1 M HF. The second anion exchange chromatography step[30,33] quantitatively removes high field strength elements (HFSE; Ti, Zr, Hf, Nb) from the W cuts. The samples were evaporated at 200 °C with added HClO$_4$ to destroy organic compounds, re-dissolved in 0.6 M HF–0.2% H$_2$O$_2$ and loaded onto pre-cleaned BioRad Polyprep columns filled with 1 ml anion exchange resin (BioRad AG1 × 8, 200–400 mesh). Titanium, Zr and Hf were rinsed off the column with 10 ml 1 M HCl–2% H$_2$O$_2$, followed by 10 ml 8 M HCl–0.01 M HF. Finally, W was eluted in 8.5 ml 6 M HCl–1 M HF. The W cuts were evaporated to dryness with added HClO$_4$ (200 °C), converted, and re-dissolved in 0.56 M HNO$_3$–0.24 M HF measurement solutions. The chemical separation of Hf and W for the spiked aliquots was accomplished using ion exchange chromatography techniques described previously[31]. Total procedural blanks were $\sim$50–300 pg for the W isotope composition analyses, and insignificant given the amounts of W in our samples (W$_{sample}$/W$_{blank}$ ≈ 250–2,000).

Procedures for W isotope measurements by MC-ICPMS. The W isotope measurements were performed using a ThermoScientific Neptune Plus MC-ICPMS in the Institut für Planetologie at the University of Münster. The analytical protocol for high-precision W isotope analyses is based on previously reported procedures[12,33,34]. Samples and standards for W isotope analyses were introduced using an ESI self-aspirating PFA nebulizer (50–60 µl min^{-1}) connected to a Cetac Aridus II desolvator system. Measurements were performed in low-resolution mode using Jet sampler and X-skimmer cones. All four major W isotopes (^{182}W, ^{183}W, ^{184}W, ^{186}W) were measured simultaneously. Total ion beams of $\sim$(2.0–3.2) × 10^{-10} were obtained for a $\sim$30 p.p.b. W standard solution at an uptake rate of $\sim$60 µl min^{-1}. Electronic baselines were obtained before each sample measurement by deflecting the beam using the electrostatic analyser for 60 s and then subtracted from sample signal intensities. A single W isotope measurement comprised 200 cycles of 4.2 s integration time each. Most samples were measured at least twice, or more often depending on the amount of W that was available for analysis (Extended Data Table 2). Small isobaric interferences from ^{184}Os and ^{186}Os on W isotope ratios were corrected by monitoring interference-free ^{188}Os, and were generally much smaller than 5 p.p.m. Only one sample (68115, BG04) required a slightly larger interference correction of $\sim$15 p.p.m. on ε^{182}W (6/4). Instrumental mass bias was corrected by normalization to either ^{186}W/^{183}W = 1.9859 (denoted '6/3') or ^{186}W/^{184}W = 0.92767 (denoted '6/4') using the exponential law. The W isotope analyses of samples were bracketed by measurements of terrestrial solution standards (Alfa Aesar) and results are reported as ε-unit (that is, 0.01%) deviations from the mean values of the bracketing standards, whose concentrations match those of the sample solutions to

within less than $\sim$20%. The reported ε^iW in Extended Data Table 2 represent the mean of pooled solution replicates (N = 1–5) together with their associated external uncertainties (see below).

Accuracy and precision of W isotope measurements. The accuracy and reproducibility of our analytical routine were assessed by repeated analyses of terrestrial rock standards (BHVO-2, BCR-2, AGV-2) whose W concentrations were similar to, or slightly lower than, those of the investigated lunar samples. In each analytical session, separate digestions of these standards were processed through the full chemical separation procedure and analysed alongside the lunar samples. The measurements of the terrestrial standards yield precise ε^{182}W values that are indistinguishable from the value of the terrestrial solution standard (Extended Data Fig. 1, Extended Data Table 1). Reported uncertainties on ε^{182}W of lunar samples are based either on the 2 s.d. obtained for terrestrial standards (in the case $N < 4$), or on the 95% confidence interval of the mean obtained for pooled solution replicates of sample analyses (that is, according to (s.d. $\times$ $t_{95\% \text{ conf.}, N-1}$)/$\sqrt{N}$). The latter is justified given that the mean ε^{182}W (6/4) obtained for the terrestrial rock standards yields ε^{182}W (6/4) = 0.00 ± 0.02 (95% conf., N = 37), demonstrating the high level of accuracy of the W isotope measurements. Some standard (and sample) analyses showed small anomalies for normalizations involving ^{183}W, including excesses in $\varepsilon^{182/183}$W (6/3) (up to +0.22) and in ε^{184}W (6/3) (up to +0.13), and deficits in ε^{183}W (6/4) (up to −0.19) (Extended Data Tables 1, 2). These coupled ε^{182}W−ε^{183}W systematics have previously been observed in high-precision MC-ICPMS studies for terrestrial standards as well as for silicate rock and iron meteorite samples[5,12,32–34], and are attributed to a mass-independent W isotope fractionation that exclusively affects ^{183}W and that is thought to be induced by W-loss during re-dissolution of the samples in Savillex beakers. The measured $\varepsilon^{182/183}$W (6/3) compositions of terrestrial standards and lunar samples can be corrected for the analytical effect on ^{183}W using different normalization schemes for W isotope measurements and the results obtained for the terrestrial standard[5,12,32–34], according to $\varepsilon^{182/183}$W (6/3)$_{corr.}$ = $\varepsilon^{182/183}$W (6/3)$_{meas.}$ − 2 × ε^{184}W (6/3)$_{meas.}$. The corrected $\varepsilon^{182/183}$W (6/3) values are indistinguishable from and as precise as the measured $\varepsilon^{182/184}$W (6/4), demonstrating that the corrections are accurate (Extended Data Tables 1-2). Note that ε^{182}W (6/4), that is, the value we use throughout this study, is not affected by the ^{183}W-effect. Thus, the ^{183}W-effect does not compromise any of the conclusions drawn from the W isotope data.

HSE concentrations by isotope dilution. The HSE concentrations of three samples (14321, 68115 and 68815) were determined by isotope dilution. The analytical techniques for HSE analyses are based on procedures described elsewhere[16,35,36], and are briefly summarized below. Samples derive from the same sample powders as used for W isotope composition analyses. Approximately 100 mg powdered sample material was weighed into pre-cleaned Carius tubes. Mixed ^{99}Ru–^{105}Pd–^{191}Ir–^{194}Pt and ^{185}Re–^{190}Os spike solutions were subsequently added followed by 1 ml concentrated HCl and 2 ml concentrated HNO$_3$. Digestion of the samples in the Carius tubes was performed in an oven at 230 °C for 48 h. Osmium was subsequently separated from the sample solutions by solvent extraction into CCl$_4$ and then back-extracted into HBr[37]. Rhenium, Ru, Pt, Ir and Pd were separated from the sample matrix by cation exchange chromatography (10 ml BioRad 50W-X8, 100–200 mesh). Samples were loaded onto the column in 2.5 ml 0.2 M HCl. The HSE were then eluted immediately in a subsequent rinse with 14 ml 0.2 M HCl. The sample solutions were evaporated to dryness and re-dissolved in 2.5 ml 0.28 M HNO$_3$ running solutions. All HSE ratios were measured using a ThermoScientific X-series II quadrupole ICPMS at the Institut für Planetologie. Total procedural blanks were $\sim$0.2 pg for Re, <0.1 pg for Os, $\sim$2 pg for Ir, $\sim$4 pg for Ru, $\sim$35 pg for Pt, and $\sim$15 pg for Pd. The corresponding blank corrections were <0.2% for Re, <0.1% for Os, <0.2% for Ir, <0.2% for Ru, <1.5% for Pt and <0.8% for Pd. As the uncertainties of HSE abundances mainly derive from the blank corrections, the external reproducibility was estimated from the difference of maximum and minimum blank corrections, and assuming an uncertainty on the blank of 50%. This yields uncertainties that are better than 1.5% (2σ) for lunar sample analyses.

Pre-exposure ε^{182}W of the bulk silicate Moon. The crystallization of a lunar magma ocean (LMO)[38–40] produced reservoirs having distinct Hf/W ratios[10,11,41,42], but lunar differentiation at $\sim$4.4 Ga (refs 17, 18) occurred too late (that is, after ^{182}Hf extinction) to generate radiogenic ε^{182}W variations within the Moon. This is consistent with the ε^{182}W values of metals from low- and high-Ti mare basalts[10] having no ε^{180}Hf anomalies[13] and, hence, no neutron capture effects (12004: ε^{182}W = 0.05 ± 0.50 (2 s.d.); 74255: ε^{182}W = 0.11 ± 0.40 (2 s.d.); see ref. 10), which are indistinguishable from the KREEP value determined here. Mare basalts have very low HSE contents[3,14], indicating minor meteorite contamination with no significant effect on ε^{182}W. Thus, there is no evidence that the pre-exposure ε^{182}W of the mare basalts would be different from that of KREEP. We, therefore, interpret

the more precise pre-exposure $\varepsilon^{182}W = 0.27 \pm 0.04$ of KREEP to represent that of the bulk silicate Moon.

Effect of meteorite contamination on $\varepsilon^{182}W$. The primordial lunar crust typically is largely devoid of HSE and elevated HSE abundances in some lunar samples are interpreted to reflect contamination with meteoritic impact components at the lunar surface[43–47]. As most meteorites have $\varepsilon^{182}W$ values markedly below those of lunar rocks, any meteoritic component can potentially shift the $\varepsilon^{182}W$ of a lunar sample to lower values. Here we quantify the magnitude of such meteoritic contamination on $\varepsilon^{182}W$ for three key samples that have no resolvable Hf isotope anomaly (14321, 68115, 68815). The HSE concentrations of these three samples were determined by isotope dilution (see Methods) and are given in Extended Data Table 3. The CI-chondrite- and Ir-normalized HSE concentrations of the three samples are shown in Extended Data Fig. 2. The three samples show elevated HSE contents and fractionated HSE patterns with suprachondritic ratios of Ru/Ir, Pt/Ir, Pd/Ir and Re/Os and slightly subchondritic Os/Ir. Such HSE patterns have previously been documented for Apollo 16 impact-melt rocks and probably reflect a mixture of IVA iron-meteorite-like (30%) and carbonaceous chondrite-like (70%) impactor components[47]. Here we assume that this mixture of impactor components had the following composition: $[Ir] = 751$ p.p.b., $[W] = 234$ p.p.b., and a $\varepsilon^{182}W$ value of -2.38 (ref. 12, 47, 48). Corrections of measured $\varepsilon^{182}W$ for meteoritic contamination were made using this composition and the measured W and Ir concentrations of the samples. The corrections on $\varepsilon^{182}W$ are ~0.04 and ~0.03 for samples 68115 and 68815 (Table 1). Because of its high W concentration (1,902 p.p.b.; Extended Data Table 3), no correction was needed for sample 14321. Its measured $\varepsilon^{182}W$ of $+0.27 \pm 0.04$ (95% confidence interval) thus defines the pre-exposure $\varepsilon^{182}W$ of KREEP directly. For all other samples, the HSE abundances could not be measured directly and previously reported HSE concentration data (if available) were used instead. For these samples the corrections were less than ~0.02 and, hence, insignificant.

Composition of the late veneer. Information about the mass and composition of the late veneer is provided by the relative and absolute HSE abundances and ratios of Se, Te, and S in Earth's primitive upper mantle[2,8]. Although HSE abundances in Earth's mantle are roughly chondritic, some HSE ratios appear to be slightly fractionated. In particular, compared to chondrites the Earth's mantle is characterized by marked excesses in Pd and Ru relative to Pt, Ir and Os (refs 20, 35, 49, 50). Moreover, the $^{187}Os/^{188}Os$ composition of the primitive upper mantle differs from that of carbonaceous chondrites, but is similar to ordinary and enstatite chondrites[2,35,36,51]. In contrast, the Se/Te ratio of the Earth's mantle is similar to that of carbonaceous chondrites, but different from any other group of chondrites[19]. These contrasting observations as well as the fractionated HSE pattern of the Earth's mantle can be reconciled if the late veneer consisted of carbonaceous chondrites with a minor fraction of differentiated, iron-meteorite-like material[35,36,47]. Here we adopt a mixture of carbonaceous (CI+CM+CV) chondrite material (80%) with an added differentiated IVA-like iron-meteorite component (20%) to represent the composition of the late veneer as previously suggested for late-accreted material on the Moon[47]. This late veneer composition corresponds to a late veneer mass of $\sim0.35\%$ Earth masses. Alternative late veneer compositions may explain some but not all of the aforementioned HSE and chalcogen signatures. For example, an ordinary-chondrite-like late veneer would explain the $^{187}Os/^{188}Os$ of the BSE, but not its suprachondritic Ru/Ir and Pd/Ir and its Se-Te systematics. Similarly, although a purely carbonaceous-chondrite-like late veneer can explain the Se–Te systematics, it cannot account for the elevated Ru/Ir, Pd/Ir, and the $^{187}Os/^{188}Os$ ratio of the BSE. We, therefore, consider a mixture of carbonaceous-chondrite-like (80%) and iron-meteorite-like (20%) material to be the current best estimate for the composition of the late veneer.

Effect of the late veneer on $\varepsilon^{182}W$ of the BSE. As chondrites and iron meteorites show deficits in $\varepsilon^{182}W$ relative to the BSE, a late veneer with the composition inferred above would have added ^{182}W-depleted material to the BSE. Mass balance calculations indicate that the late veneer lowered the $\varepsilon^{182}W$ of the pre-late-veneer BSE by 0.16–0.38 (Extended Data Table 4, Fig. 1). The effect of the late veneer on $\varepsilon^{182}W$ of the BSE does not change significantly for alternative assumed late veneer compositions and proportions (Extended Data Table 4). Note that the effect is slightly larger for our preferred composition of a mixture of carbonaceous chondrite and iron meteorite than for a late veneer composed entirely of known groups of chondrites, because the presence of W-rich and ^{182}W-depleted iron meteorite material results in a higher W concentration of ~200 p.p.b. and a more negative $\varepsilon^{182}W$ of about -2.6 for the late veneer (using W concentrations of ~113 p.p.b. for CI chondrites[31], ~130 p.p.b. for CM chondrites[31], ~170 p.p.b. for CV chondrites[31], and 480 p.p.b. for the IVA iron meteorites[48], and $\varepsilon^{182}W$ values of -2.2 for CI chondrites, -1.75 for CM chondrites, -2.0 for CV chondrites and -3.3 for IVA iron meteorites[12,31]). For the present-day BSE we used a W concentration of 13 ± 5 p.p.b. (2σ), which is based on the W/U ratio of the BSE of 0.64 ± 0.05 (2 s.e., $N = 86$)[52] and assuming a U concentration for the BSE of 20 ± 8 p.p.b.

(2σ)[53–55]. The W concentration of the BSE of 13 ± 5 p.p.b. (2σ) used here is in good agreement with an independent recent estimate of ~12 p.p.b. (ref. 56). The $\varepsilon^{182}W$ of the present-day BSE is 0 by definition.

Effect of giant impact on $\varepsilon^{182}W$ of BSE and Moon. The W isotope compositions of both the pre-late-veneer BSE and the Moon reflect specific mixtures of the same three components: (1) proto-Earth mantle, (2) impactor mantle, and (3) impactor core. Using simple mass balance calculations we demonstrate that the giant impact most probably should have produced distinct $\varepsilon^{182}W$ of the pre-late-veneer BSE and the Moon, for two main reasons. First, the giant impact would have modified the $\varepsilon^{182}W$ of the proto-Earth's mantle, but the material forming the Moon would most probably have been ejected earlier, that is, before the impactor material could have equilibrated fully with the Earth's mantle[1,26,28]. Second, some impactor core material should be present in the lunar accretion disk, and this core material would have had a strong effect on the $\varepsilon^{182}W$ of the bulk Moon. Extended Data Fig. 3 shows the effects of the giant impact on $\varepsilon^{182}W$ of the BSE, plotted as a function of the mass fraction of the impactor relative to Earth's mass $(M_{\mathrm{Imp}}/M_{\oplus})$. Shown are the effects for two different impactor compositions, Mars-like or Vesta-like. In terms of their Hf–W systematics, Mars (low mantle Hf/W and $\varepsilon^{182}W$) and Vesta (high mantle Hf/W and radiogenic $\varepsilon^{182}W$) may be considered end members among (known) differentiated planetary bodies in the inner Solar System. The $\varepsilon^{182}W$ of the proto-Earth's mantle is set to zero in the calculations, so that the calculated effects show the difference in $\varepsilon^{182}W$ between the pre- and post-giant impact BSE. Note that assuming a Mars-like impactor represents the end member case of very similar $\varepsilon^{182}W$ in impactor and proto-Earth's mantle, whereas assuming a Vesta-like impactor is the end member case of very different $\varepsilon^{182}W$ in impactor and proto-Earth's mantle. While material from the impactor mantle is directly added to the proto-Earth's mantle, impactor core material is divided between a fraction k that chemically equilibrates with the proto-Earth's mantle and a fraction $(1 - k)$ that directly joins Earth's core without any prior equilibration. Our calculations show that in most cases the giant impact induces a significant shift in the $\varepsilon^{182}W$ value of the BSE, that is, the pre- and post-giant impact $\varepsilon^{182}W$ values of the BSE are probably different (Extended Data Fig. 3). Consequently, even if the Moon predominantly consists of proto-Earth mantle material—as in some recent models of the giant impact[26,27,28]—an $\varepsilon^{182}W$ difference between the BSE and the Moon would be expected. The smallest giant-impact-induced $\varepsilon^{182}W$ shifts are observed for a Mars-like impactor and small impactor-to-Earth ratios. However, even in this case the shift of the BSE $\varepsilon^{182}W$ value would only be small for very low values of k, that is, for low degrees of equilibration of the impactor core with proto-Earth's mantle. For small impactors the degree of equilibration is expected to be high, however[57]. Thus, even the addition of a small impactor having an $\varepsilon^{182}W$ value similar to that of the proto-Earth's mantle would probably induce a significant shift in the $\varepsilon^{182}W$ value of the BSE. Nevertheless, our calculations show that regardless of the $\varepsilon^{182}W$ of the impactor mantle, a specific combination of impactor composition and degree of equilibration of impactor with proto-Earth's mantle can result in no measurable change of $\varepsilon^{182}W$ in the BSE (Extended Data Fig. 3). However, such specific combinations are low-probability cases and it seems far more likely that the giant impact significantly modified the $\varepsilon^{182}W$ of the proto-Earth's mantle.

Another reason why the ^{182}W similarity between the Moon and the pre-late-veneer BSE is remarkable relates to the prediction of most giant-impact simulations that a small fraction of impactor core material is present in the lunar accretion disk[1,26,28]. This material can also account for the presence of a small lunar core[58,59], which otherwise might be difficult to explain. Owing to its high W content and presumably unradiogenic $\varepsilon^{182}W$, the addition of even minute amounts of impactor core material will significantly lower the $\varepsilon^{182}W$ of the proto-lunar material (Extended Data Fig. 4). For a possible range of impactor core compositions and assuming 2.5% impactor core material in the lunar disk (that is, equivalent to the estimated size of the lunar core), the $\varepsilon^{182}W$ of the proto-lunar material would have changed by at least 0.7 but more probably by several ε-units (Extended Data Fig. 4). This effect is large compared to the very close agreement of the pre-late-veneer $\varepsilon^{182}W$ of the BSE and the $\varepsilon^{182}W$ of the Moon.

In summary, both the presence of impactor core material within the lunar accretion disk as well as the effects of the giant impact on the $\varepsilon^{182}W$ of the proto-Earth's mantle are likely to have produced ^{182}W variations in the ε-unit range, yet the $\varepsilon^{182}W$ value of the pre-late-veneer BSE and our newly defined $\varepsilon^{182}W$ value for the Moon are similar to within ~10 p.p.m. It remains possible to explain this similarity by a very specific combination of several parameters[25], including (1) specific compositions of impactor mantle, impactor core and proto-Earth's mantle, (2) degree of re-equilibration of impactor core material within the proto-Earth's mantle, and (3) a particular fraction of impactor core material in the lunar accretion disk. Yet, this would imply that the high degree of similarity in $\varepsilon^{182}W$ of the pre-late-veneer BSE and the Moon is purely coincidental, which seems highly unlikely.

Sample size. No statistical methods were used to predetermine sample size.

30. Kleine, T., Hans, U., Irving, A. J. & Bourdon, B. Chronology of the angrite parent body and implications for core formation in protoplanets. *Geochim. Cosmochim. Acta* **84,** 186–203 (2012).
31. Kleine, T., Mezger, K., Münker, C., Palme, H. & Bischoff, A. ^{182}Hf-^{182}W isotope systematics of chondrites, eucrites, and martian meteorites: chronology of core formation and early mantle differentiation in Vesta and Mars. *Geochim. Cosmochim. Acta* **68,** 2935–2946 (2004).
32. Kruijer, T. S. *et al.* Hf–W chronometry of core formation in planetesimals inferred from weakly irradiated iron meteorites. *Geochim. Cosmochim. Acta* **99,** 287–304 (2012).
33. Kruijer, T. S., Kleine, T., Fischer-Gödde, M., Burkhardt, C. & Wieler, R. Nucleosynthetic W isotope anomalies and the Hf-W chronometry of Ca-Al-rich inclusions. *Earth Planet. Sci. Lett.* **403,** 317–327 (2014).
34. Kruijer, T. S. *et al.* Neutron capture on Pt isotopes in iron meteorites and the Hf–W chronology of core formation in planetesimals. *Earth Planet. Sci. Lett.* **361,** 162–172 (2013).
35. Fischer-Gödde, M., Becker, H. & Wombacher, F. Rhodium, gold and other highly siderophile elements in orogenic peridotites and peridotite xenoliths. *Chem. Geol.* **280,** 365–383 (2011).
36. Fischer-Gödde, M., Becker, H. & Wombacher, F. Rhodium, gold and other highly siderophile element abundances in chondritic meteorites. *Geochim. Cosmochim. Acta* **74,** 356–379 (2010).
37. Cohen, A. S. & Waters, F. G. Separation of osmium from geological materials by solvent extraction for analysis by thermal ionisation mass spectrometry. *Anal. Chim. Acta* **332,** 269–275 (1996).
38. Wood, J. A., Dickey, J. S., Marvin, U. B. & Powell, B. N. Lunar anorthosites and a geophysical model of the Moon. *Proc. Apollo 11 Lunar Sci. Conf.* **1,** 965–988 (1970).
39. Taylor, S. R. & Jakeš, P. The geochemical evolution of the Moon. *Proc. Fifth Lunar Conf.* **2,** 1287–1305 (1974).
40. Warren, P. H. & Wasson, J. T. The origin of KREEP. *Rev. Geophys.* **17,** 73–88 (1979).
41. Righter, K. & Shearer, C. K. Magmatic fractionation of Hf and W: constraints on the timing of core formation and differentiation in the Moon and Mars. *Geochim. Cosmochim. Acta* **67,** 2497–2507 (2003).
42. Münker, C. A high field strength element perspective on early lunar differentiation. *Geochim. Cosmochim. Acta* **74,** 7340–7361 (2010).
43. Morgan, J. W., Walker, R. J., Brandon, A. D. & Horan, M. F. Siderophile elements in Earth's upper mantle and lunar breccias: data synthesis suggests manifestations of the same late influx. *Meteorit. Planet. Sci.* **36,** 1257–1275 (2001).
44. Korotev, R. L. The meteorite component of Apollo 16 noritic impact melt breccias. *J. Geophys. Res.* **92,** E491 (1987).
45. Puchtel, I. S., Walker, R. J., James, O. B. & Kring, D. A. Osmium isotope and highly siderophile element systematics of lunar impact melt breccias: implications for the late accretion history of the Moon and Earth. *Geochim. Cosmochim. Acta* **72,** 3022–3042 (2008).
46. Norman, M. D., Bennett, V. C. & Ryder, G. Targeting the impactors?: siderophile element signatures of lunar impact melts from Serenitatis. *Geochim. Cosmochim. Acta* **202,** 217–228 (2002).
47. Fischer-Gödde, M. & Becker, H. Osmium isotope and highly siderophile element constraints on ages and nature of meteoritic components in ancient lunar impact rocks. *Geochim. Cosmochim. Acta* **77,** 135–156 (2012).
48. McCoy, T. J. *et al.* Group IVA irons: new constraints on the crystallization and cooling history of an asteroidal core with a complex history. *Geochim. Cosmochim. Acta* **75,** 6821–6843 (2011).
49. Pattou, L., Lorand, J. P. & Gros, M. Non-chondritic platinum-group element ratios in the Earth's mantle. *Nature* **379,** 712–715 (1996).
50. Horan, M. Highly siderophile elements in chondrites. *Chem. Geol.* **196,** 27–42 (2003).
51. Meisel, T., Walker, R. J., Irving, A. J. & Lorand, J.-P. Osmium isotopic compositions of mantle xenoliths: a global perspective. *Geochim. Cosmochim. Acta* **65,** 1311–1323 (2001).
52. Arevalo, R. & McDonough, W. F. Tungsten geochemistry and implications for understanding the Earth's interior. *Earth Planet. Sci. Lett.* **272,** 656–665 (2008).
53. McDonough, W. F. & Sun, S.-s. The composition of the Earth. *Chem. Geol.* **120,** 223–253 (1995).
54. Lyubetskaya, T. & Korenaga, J. Chemical composition of Earth's primitive mantle and its variance: 1. Method and results. *J. Geophys. Res.* **112,** B03211 (2007).
55. Palme, H. & O'Neill, H. S. C. *Cosmochemical Estimates of Mantle Composition. Treatise on Geochemistry* (2nd edn) **3,** 1–39 (Elsevier, 2014).
56. König, S. *et al.* The Earth's tungsten budget during mantle melting and crust formation. *Geochim. Cosmochim. Acta* **75,** 2119–2136 (2011).
57. Dahl, T. W. & Stevenson, D. J. Turbulent mixing of metal and silicate during planet accretion — And interpretation of the Hf-W chronometer. *Earth Planet. Sci. Lett.* **295,** 177–186 (2010).
58. Khan, A., Maclennan, J., Taylor, S. R. & Connolly, J. A. D. Are the Earth and the Moon compositionally alike? Inferences on lunar composition and implications for lunar origin and evolution from geophysical modeling. *J. Geophys. Res.* **111,** E05005 (2006).
59. Weber, R. C., Lin, P.-Y., Garnero, E. J., Williams, Q. & Lognonné, P. Seismic detection of the lunar core. *Science* **331,** 309–312 (2011).
60. Palme, H. Lodders, K. & Jones, A. *Solar System Abundances of the Elements. Treatise on Geochemistry* (2nd edn), **2,** 15–36 (Elsevier, 2014).
61. Cottrell, E., Walter, M. J. & Walker, D. Metal–silicate partitioning of tungsten at high pressure and temperature: Implications for equilibrium core formation in Earth. *Earth Planet. Sci. Lett.* **281,** 275–287 (2009).

LETTER

Eocene primates of South America and the African origins of New World monkeys

Mariano Bond[1], Marcelo F. Tejedor[2,3], Kenneth E. Campbell Jr[4], Laura Chornogubsky[5,6], Nelson Novo[2] & Francisco Goin[1]

The platyrrhine primates, or New World monkeys, are immigrant mammals whose fossil record comes from Tertiary and Quaternary sediments of South America and the Caribbean Greater Antilles[1,2]. The time and place of platyrrhine origins are some of the most controversial issues in primate palaeontology, although an African Palaeogene ancestry has been presumed by most primatologists[3,4]. Until now, the oldest fossil records of New World monkeys have come from Salla, Bolivia[5,6], and date to approximately 26 million years ago[7], or the Late Oligocene epoch. Here we report the discovery of new primates from the ?Late Eocene epoch of Amazonian Peru, which extends the fossil record of primates in South America back approximately 10 million years. The new specimens are important for understanding the origin and early evolution of modern platyrrhine primates because they bear little resemblance to any extinct or living South American primate, but they do bear striking resemblances to Eocene African anthropoids, and our phylogenetic analysis suggests a relationship with African taxa. The discovery of these new primates brings the first appearance datum of caviomorph rodents and primates in South America back into close correspondence, but raises new questions about the timing and means of arrival of these two mammalian groups.

The primate material reported on here comprises one complete and two incomplete upper molars and one complete lower molar from the Santa Rosa local fauna of eastern Peru[8] (Extended Data Fig. 1), all of which have a generalized crown morphology. The upper molars have vague resemblances to small callitrichines (living marmosets and tamarins), but with primitive morphologies closer to Palaeogene African anthropoids. The lower molar also exhibits an undoubted primitive pattern resembling some Palaeogene African anthropoids. Thus, not only do the Santa Rosa specimens represent the oldest New World monkeys known and hint strongly towards an African ancestry, they also alter prior perceptions and interpretations of early New World monkey morphotypes.

<div align="center">

Order Primates Linnaeus, 1758
Suborder Anthropoidea Mivart, 1864
Parvorder ?Platyrrhini E. Geoffroy, 1812
Family *incertae sedis*
Perupithecus gen. nov.

</div>

Generic diagnosis. *Perupithecus* differs from all platyrrhines, except the extinct *Branisella*, *Szalatavus*, *Lagonimico*, and *Micodon* and the extant Callitrichinae by having an upper first molar (M[?1]) with subtriangular outline with a narrow lingual side. The M[?1] differs from those of all known platyrrhines in having paracone, metacone, and protocone connected by more trenchant crests surrounding a moderately deep trigon basin. Trigon reduction in *Perupithecus* M[?1] is compensated by a more expanded side lingual to protocone, thus differing from all platyrrhines, which have a more enlarged trigon and less developed lingual shelf. An entoprotocrista (absent from first upper molars in all

known platyrrhines) descends from tip of protocone to reach the lingual cingulum, isolating a precingulum that extends mesiolingually to the anterior face of the protocone. The crestiform hypocone of *Perupithecus* M[?1] resembles that of some Callitrichinae, but differs from those in Pitheciidae, Atelidae, and Cebinae, where the upper molars are more quadrangular with developed hypocones. The M[?1] has a deep mesial fovea formed by a preprotocrista that divides buccally into a preparaconular crista that reaches the parastyle and a hypoparacrista that contacts the paracone (mesial fovea less developed and shallower in *Branisella*, reduced or absent in Callitrichinae, and occasionally present, but shallower, in other platyrrhines).

Type species. *Perupithecus ucayaliensis*.

Etymology. *Peru*, from Peru, the country of origin, and *pithecus*, the Greek word for monkey.

<div align="center">

Perupithecus ucayaliensis sp. nov.

</div>

Figures 1–3a and Extended Data Fig. 2.

Holotype. Complete upper left molar (M[?1]), Instituto Geológico, Minero y Metalúrgico (INGEMMET), Lima, Collección Paleontológica del INGEMMET (CPI) 6486. Natural History Museum of Los Angeles County (LACM) 6289/155085; plastotype.

Etymology. *ucayaliensis*, after the department in which the Santa Rosa locality occurs.

Type locality and age. Santa Rosa, Atalaya Province, Ucayali Department, Peru. Left bank of Yurúa River, at 9° 29′ 39′ S, 72° 45′ 48′ W. Yahuarango Formation. LACM locality 6289 (ref. 9). ?Late Eocene (see discussion of locality and age in Supplementary Discussion 1).

Species diagnosis. Same as for genus.

Measurements. Labiolingual, 3.5 mm; mesiodistal, 2.6 mm.

Description. *Perupithecus ucayaliensis* was a small primate, comparable in size to the tamarins *Callimico* and some species of *Saguinus*, among
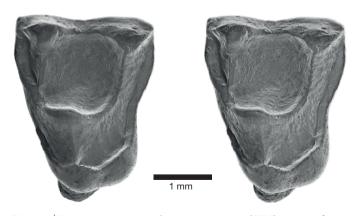
Figure 1 | Stereo-pair scanning electron microscope (SEM) images of *Perupithecus ucayaliensis*, holotype. Left M[?1], CPI-6486 (LACM 6289/155085), in occlusal view. Scale bar, 1 mm.

[1]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), División Paleontología Vertebrados, Museo de Ciencias Naturales de La Plata, B1900FWA La Plata, Argentina. [2]CONICET, Centro Nacional Patagónico, Boulevard Almirante Brown 2915, 9120 Puerto Madryn, Chubut, Argentina. [3]Facultad de Ciencias Naturales, Sede Trelew, Universidad Nacional de la Patagonia 'San Juan Bosco', 9100 Trelew, Chubut, Argentina. [4]Vertebrate Zoology, Natural History Museum of Los Angeles County, Los Angeles, California 90007, USA. [5]CONICET, Sección Paleontología de Vertebrados, Museo Argentino de Ciencias Naturales 'Bernardino Rivadavia', Avenida Ángel Gallardo 470, C1405DJR Buenos Aires, Argentina. [6]Universidad Nacional de Luján, Departamento de Ciencias Básicas. Ruta Nacional 5 y Avenida Constitución, 6700 Luján, Buenos Aires Province, Argentina.
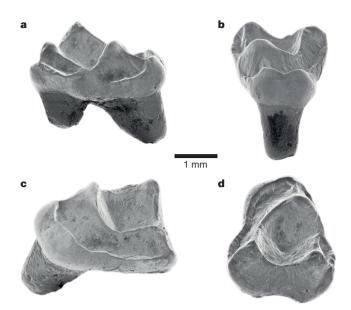
Figure 2 | SEM images of *Perupithecus ucayaliensis*, holotype. CPI-6486 (LACM 6289/155085) (left M$^{?1}$), in (**a**) mesiolingual, (**b**) lingual, (**c**) distolingual, and (**d**) occlusobuccal views. Note the deep mesial fovea and entoprotocrista in **a**, the crestiform hypocone in **b**, the proportions between the trigon basin and lingual side in **c**, and the flat buccal side in **d**, as discussed in the text. Scale bar, 1 mm.

the callitrichines. Although the upper molar of *Perupithecus* is distinct from that of all known platyrrhines, it does resemble some extinct Eocene African anthropoids, such as *Talahpithecus*[10] (Fig. 3b) and *Proteopithecus*[11], by having M$^{?1}$ with subtriangular outline with a narrow lingual side hosting an entoprotocrista. The M$^{?1}$ further resembles that of *Talahpithecus* by having a crestiform hypocone, an expanded side lingual to protocone, and protocone connected to paracone and metacone by trenchant crests surrounding the trigon.

The holotype, INGEMMET CPI-6486 (LACM 6289/155085; plastotype), has been interpreted as a left M$^{?1}$ primarily because the paracone and metacone are aligned buccally (Fig. 1). The metacone is placed more lingually in first upper molars of *Branisella* and callitrichines. The subtriangular crown of *Perupithecus* has three main cusps of comparable size connected by trenchant crests completely enclosing a moderately deep trigon. The cusps have little apical wear, with the metacone worn the most. Buccally, the paracone and metacone are connected by descending postparacrista and premetacrista, forming a 'V' shape in lateral view. Parastyle and metastyle are present, as in *Branisella*, *Szalatavus*, and unworn upper molars of some callitrichines, as well as in *Talahpithecus*, although the metastyle is poorly developed in the latter, as in *Perupithecus*. There is no mesostyle.

The protocone, paracone, and metacone are well-developed, but not basally enlarged as seen in *Branisella* and *Szalatavus*, with the protocone slightly higher and tilted buccally. The cusps and entire crown are lower than observed in *Branisella* and *Szalatavus*. The preprotocrista and

postprotocrista meet the preparacrista and hypometacrista, respectively, in the middle part of the trigon. The preprotocrista ends in a deep mesial fovea in an area with a small bulge similar to a vestigial paraconular area. *Talahpithecus* also has a well-developed, but less deep, mesial fovea.

There is a small swelling (vestigial entostyle or pericone?) where the entoprotocrista contacts the lingual cingulum. The lingual cingulum connects distally to a crestiform hypocone with a slightly enlarged base, as seen in *Callimico*. Posterior to the hypocone, the postcingulum is complete, reaching the metastylar area. There is no talon enlargement, which gives a subtriangular outline similar to that of *Branisella*, *Szalatavus*, and callitrichines, as well as some African taxa such as *Talahpithecus* and *Proteopithecus*. The molar has one well-developed lingual root and two buccal roots. The molar crown morphology of *Perupithecus* is consistent with an insectivorous-frugivorous diet.
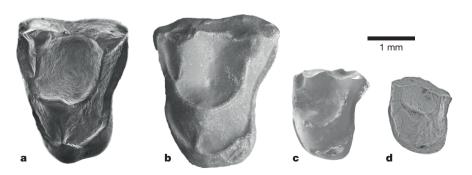
Gen. et sp. *incertae sedis* A

Two additional primate upper molars, a left and right M^1 or M^2 (CPI-7000 and CPI-7001, respectively; LACM 6289/157193 and 6289/157063, casts) are recorded from Santa Rosa. Both specimens (Fig. 3c, d) lack their buccal sides, and CPI-7001 also has the area encompassing the mesial fovea broken and the preprotocrista worn. The remaining morphology and smaller size (~70% of the size of *Perupithecus*) of these specimens indicate that they belong to a genus distinct from *Perupithecus*. We refrain from formally naming these specimens because their incomplete nature prevents determination of the state of several diagnostic characters.

These molars differ from *Perupithecus* in having a poorly developed mesiolingual cingulum that is interrupted lingual to the protocone before continuing towards the hypocone; a better developed hypocone placed on a distolingual expansion of the cingulum, and by lacking an entoprotocrista and its swelling on the lingual cingulum. The paracone and metacone development cannot be seen because of the broken buccal sides. A mesial fovea is present, shallower than in *Perupithecus*; and the hypoparacrista descends more vertically from the paracone to meet the postparaconule crista, whereas *Perupithecus* has a longer, less inclined hypoparacrista. The paraconule area is more developed than in *Perupithecus*. The distal part of the CPI-7000, although broken, exhibits a postcingulum that curves back behind the metacone area, giving a 'waisted' appearance, as in callitrichines, *Branisella*, *Szalatavus*, and *Perupithecus*, as well as *Talahpithecus* and other African anthropoids. The cusps are basally broader than in *Perupithecus*, and the trigon crests are less trenchant. In overall shape, the mesiolingual side with a weak cingulum, the distolingual expansion with the small hypocone, the waisted distal part, and the less developed mesial fovea are callitrichine-like characters, usually more common in first upper molars. These features, however, are also seen in first upper molars of *Branisella* and *Proteopithecus*, thus indicating a primitive pattern.

Gen. et sp. *incertae sedis* B

The Santa Rosa local fauna has also produced an isolated right M$_3$ (CPI-6487; LACM 6289/156981, cast) (Fig. 4) whose primate status is indicated by the overall cusp positions and general morphology. The specimen has an unusually high protoconid and metaconid, which prevents its attribution to either of the genera represented by the Santa

Figure 3 | Comparison SEM images. a, Holotype of *Perupithecus ucayaliensis* (CPI-6486 (LACM 6289/155085)); **b**, cast of holotype of *Talahpithecus parvus* (DT1-31); **c**, unassigned left upper molar (M^1 or M^2) (CPI-7000; LACM 6289/157193); and **d**, unassigned right (reversed) upper molar (M^1 or M^2) (CPI-7001; LACM 6289/157063), in occlusal views. Scale bar, 1 mm.
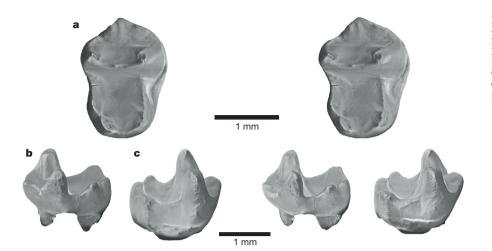
Figure 4 | Stereo-pair SEM image of unassigned primate right M$_3$. CPI-6487; LACM 6289/156981 in (**a**) occlusal, (**b**) lingual, and (**c**) buccal views. Note the twinned entoconid-hypoconulid, as well as the low paraconid, in **b**, and the reduced buccal cingulid and height of the protoconid and metaconid in **c**. Scale bar, 1 mm.

Rosa upper molars, where the trigon basins are not particularly deep and the cusps are lower and less differentiated.

This molar has five cusps, with a pointed protoconid and a lower metaconid transversely aligned with medial and lateral protocristids; the latter are connected to each other, closing the trigonid distally and separating it from the talonid. The cusps are relatively broad at the base and have an occlusal relief with moderate shearing crests (not pointed cusps and sharp crests as expected in primitive bats or insectivores). Distinctive features include a reduced paraconid in a restricted trigonid; a trigonid slightly higher relative to the talonid; a remnant buccal cingulum and no distal cingulum; a hypoconid not pointed nor buccally displaced; and a relatively developed and moderately deep talonid basin with narrow buccolingual margins with respect to the trigonid, as typical in lower third molars. A low paraconid connects to the protoconid through a paracristid descending almost vertically from the tip of the protoconid. A small sulcus separates the paraconid from the premetacristid that descends from the metaconid. The talonid has a moderate cristid obliqua that connects distally with the protoconid. The hypoconid is well-developed with a moderate posthypocristid. The twinned hypoconulid and entoconid are separated by a shallow sulcus. The entocristid is short, meeting the descending and also short postmetacristid. The trigonid is wider than the talonid and proportionally large (more than one-third the total length of the molar), and the talonid is relatively narrow, suggesting that CPI-6487 is a third lower molar.

A developed hypoconulid twinned to the entoconid is a primitive character seen in several Palaeogene anthropoids from Africa, such as *Proteopithecus*, *Talahpithecus*, and the Oligopithecidae. The paraconid placed almost mesially, which gives a triangular occlusal shape to the trigonid, also illustrates the primitiveness of this lower molar. In general, platyrrhines tend to shorten the trigonid mesiodistally, losing the paraconid, except for some individuals where the paraconid is insignificant and does not affect the buccolingually rectangular shape of the reduced trigonid. Only callitrichines and *Soriacebus* have trigonids larger mesiodistally, but in most cases the paraconid is lost. Compared with callitrichines, CPI-6487 has a better developed talonid with moderate cusps. Callitrichines not only trend towards reducing the talonid in M$_{1–2}$, they have extremely reduced entoconids and hypoconids in some taxa. They also lose third molars (except for *Callimico*). There are no known platyrrhines, living or extinct, with a comparable lower molar structure, and even for other anthropoids, the structure of this M$_3$ is notably primitive.

The evidence presented above documents a surprising diversity of primates for such an early site in South America. Previously, the lack of a fossil record for New World monkeys pre-dating the Salla primates led to significantly differing hypotheses on platyrrhine evolutionary history[2,12–15], most of which *Perupithecus* and the other Santa Rosa primates falsify.

The most striking aspect of *Perupithecus* is its strong resemblance to the Eocene African anthropoid *Talahpithecus* (Fig. 3b) from the late Middle Eocene (~38–39 million years (Myr)) of Dur At-Talah, Libya[10], although this locality has been suggested to be Late Eocene in age (~35–36 Myr)[16]. *Perupithecus*, like *Talahpithecus*, exhibits such anthropoid features as an M$^{?1}$ with a well-developed lingual cingulum and the presence of a hypocone, absence of paraconule, reduced stylar shelf with almost no buccal cingulum, reduced parastyle and metastyle, as well as shorter preparacrista and postmetacrista. *Talahpithecus* is similar to *Perupithecus* in size and overall morphology, and differs only in some minor details, such as having the buccal side basal to the metacone slightly developed.

Our phylogenetic analyses (see Supplementary Discussion 2 and Extended Data Figs 3 and 4) placed *Perupithecus* and *Talahpithecus* as sister taxa, a stem clade assigned with doubts to the Platyrrhini. We recognize that the characterization of *Perupithecus* and *Talahpithecus* as sister taxa in our phylogenetic analysis might be a result of morphological convergence, although we regard it as unlikely. *Talahpithecus* was originally attributed to the Oligopithecidae[10], which another study[17] included in the Catarrhini. However, the oligopithecid status of *Talahpithecus* is uncertain because it is based on three isolated teeth, two of which are broken. In addition, an undescribed anthropoid from the BQ-2 levels of Fayum represented by upper and lower dentitions has been mentioned, and close similarities with *Talahpithecus* were discussed[16]. This undescribed anthropoid had three upper and lower premolars, as in platyrrhines and in contrast to oligopithecids, and it might also represent a close relative of platyrrhines.

On the basis of the position of *Talahpithecus* and *Perupithecus* as sister taxa, closely linking the Peruvian and North African primates, the following preliminary hypotheses arise: (1) the Santa Rosa primates are members of a pre-platyrrhine anthropoid stock that differentiated in Africa from anthropoid ancestors; (2) they are Eocene platyrrhines of immediate African ancestry; or (3) platyrrhines originated in Africa, where they subsequently became extinct, and *Talahpithecus* is an early representative of their radiation, which possibly included *Proteopithecus* as part of a basal stock. *Perupithecus* fits well in any scenario depicting an African origin for platyrrhines, but more specimens are needed to determine whether the Santa Rosa primates are best explained by any of these hypotheses.

The advanced Late Eocene anthropoids of northern Africa were possibly preceded by the dispersal of stem anthropoids, the eosimiiforms (Eosimiidae + Afrotarsidae *sensu* ref. 18), from Asia to Africa by the Middle Eocene[18,19]. A possible morphological intermediate between African eosimiiforms and simiiforms is *Amamria*[19]. Although *Amamria* might be a basal anthropoid, some differences with respect to *Perupithecus* and expected early platyrrhine representatives are noted (that is, developed stylar shelf internalizing the paracone and metacone, better

developed lingual cingulum, absence of entoprotocrista). Similarly, the platyrrhine primates also might have originated via an east–west migration (that is, Africa to South America) of middle to Late Eocene anthropoids of still uncertain relationships, although crown anthropoids such as *Talahpithecus* were certainly involved.

South America and Africa have been separated since the early Late Cretaceous, so vicariance of primates does not appear reasonable as an explanation for their appearance in the Eocene on two continents separated by the Atlantic. Numerous studies have focused on the possibility of primates crossing the Atlantic to reach South America from Africa (for example, refs 20, 21), with rafting across the Atlantic usually considered a feasible way for how primates arrived in South America, presuming they originated in Africa[4]. It has also been proposed that perhaps more than one rafting event carried more than one primate taxon successfully across the Atlantic in the Palaeogene[22,23], suggesting a possible early African diversification before arrival in South America. This might favour the arrival of different higher clades in South America, thus allowing the diversification of more than one lineage. A similar means of arrival in South America has often been proposed for the hystricognath rodents[24], the dispersal of amphisbaenian and gekkotan lizards[25], and the Opisthocomiformes, a Neotropical group of birds (hoatzins) with weak flight capabilities and alleged African origin[26]. And, with the discovery of the Santa Rosa primates, the re-established, relatively contemporaneous first appearance datum of primates and rodents in South America leads to consideration of possible similarities of intercontinental dispersal mechanisms for the two mammalian groups. However, the derived status of Santa Rosa rodents relative to contemporaneous African rodents hints that rodents and primates might not have had simultaneous crossing episodes. Or, alternatively, the two groups had differing rates of diversification after arrival in South America.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Hartwig, W. C. & Meldrum, D. J. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 175–188 (Cambridge Univ. Press, 2002).
2. Tejedor, M. F. Sistemática, evolución y paleobiogeografía de los primates Platyrrhini. *Rev. Mus. La Plata Sec. Zool.* **20,** 20–39 (2013).
3. Fleagle, J. G. *Primate Adaptation and Evolution* 3rd edn (Academic, 2013).
4. de Queiroz, A. *The Monkey's Voyage* (Basic Books, 2014).
5. Hoffstetter, R. Un primate de l'Oligocène inférieur sud-américain: *Branisella boliviana* gen. et sp. nov. *C. R. Acad. Sci. D* **69,** 434–437 (1969).
6. Fleagle, J. G. & Tejedor, M. F. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 161–173 (Cambridge Univ. Press, 2002).
7. Kay, R. F., MacFadden, B. J., Madden, R. H., Sandeman, H. & Anaya, F. Revised age of the Salla beds, Bolivia, and its bearing on the age of the Deseadan South American land mammal 'age'. *J. Vertebr. Paleontol.* **18,** 189–199 (1998).
8. Campbell, K. E. Jr, ed. The Paleogene mammalian fauna of Santa Rosa, Amazonian Peru. *Nat. Hist. Mus. Los Angeles County Sci. Ser.* **40,** i–vi, 1–163 (2004).
9. Campbell, K. E. Jr, Frailey, C. D. & Romero-Pittman, L. The Paleogene Santa Rosa local fauna of Amazonian Perú: geographic and geologic setting. *Nat. Hist. Mus. Los Angeles County Sci. Ser.* **40,** 3–14 (2004).
10. Jaeger, J. J. *et al.* Late Middle Eocene epoch of Libya yields earliest known radiation of African anthropoids. *Nature* **467,** 1095–1098 (2010).
11. Simons, E. L. Description of two genera and species of Late Eocene Anthropoidea from Egypt. *Proc. Natl Acad. Sci. USA* **86,** 9956–9960 (1989).
12. Kay, R. F. *et al.* The anatomy of *Dolichocebus gaimanensis*, a stem platyrrhine monkey from Argentina. *J. Hum. Evol.* **54,** 323–382 (2008).
13. Rosenberger, A. L., Tejedor, M. F., Cooke, S. & Pekar, S. in *South American Primates: Testing New Theories in the Study of Primate Behavior, Ecology, and Conservation. Developments in Primatology: Progress and Prospects* (eds Garber, P. A., Estrada, A., Bicca-Marques, J. C., Heymann, E. W. & Strier, K. B.) Vol. 16, 69–113 (Springer, 2009).
14. Rosenberger, A. L. Platyrrhines, PAUP, parallelism, and the long lineage hypothesis: a reply to Kay et al. (2008). *J. Hum. Evol.* **59,** 214–217 (2010).
15. Kay, R. F. Biogeography in deep time – what do phylogenetics, geology, and paleoclimate tell us about early platyrrhine evolution? *Mol. Phylogenet. Evol. B* **82,** 358–374 (2015).
16. Seiffert, E. R. Early primate evolution in Afro-Arabia. *Evol. Anthropol.* **21,** 239–253 (2012).
17. Seiffert, E. R. *et al.* A fossil primate of uncertain affinities from the earliest Late Eocene of Egypt. *Proc. Natl Acad. Sci. USA* **107,** 9712–9717 (2010).
18. Chaimanee, Y. *et al.* Late Middle Eocene primate from Myanmar and the initial anthropoid colonization of Africa. *Proc. Natl Acad. Sci. USA* **109,** 10293–10297 (2012).
19. Marivaux, L. *et al.* A morphological intermediate between eosimiiform and simiiform primates from the Late Middle Eocene of Tunisia: macroevolutionary and paleobiogeographic implication of early anthropoids. *Am. J. Phys. Anthropol.* **154,** 387–401 (2014).
20. Houle, A. Floating islands: a mode of long-distance dispersal for small and medium-sized terrestrial vertebrates. *Divers. Distrib.* **4,** 201–216 (1998).
21. Bandoni de Oliveira, F., Molina, E. C. & Marroig, G. in *South American Primates: Comparative Perspectives in the Study of Behavior, Ecology, and Conservation. Developments in Primatology: Progress and Prospects* (eds Garber, P. A., Estrada, A., Bicca-Marques, J. C., Heymann, E. W. & Strier, K. B.) Vol. 16, 55–68 (Springer, 2009).
22. Bauer, K. & Schreiber, A. Double invasion of Tertiary island South America by ancestral New World monkeys? *Biol. J. Linn. Soc.* **60,** 1–20 (1997).
23. Takai, M., Anaya, F., Shigehara, N. & Setoguchi, T. New fossil materials of the earliest New World monkey, *Branisella boliviana*, and the problem of platyrrhine origins. *Am. J. Phys. Anthropol.* **111,** 263–281 (2000).
24. Antoine, P. O. *et al.* Middle Eocene rodents from Peruvian Amazonia reveal the pattern and timing of caviomorph origins and biogeography. *Proc. R. Soc. B* **279,** 1319–1326 (2011).
25. Gamble, T. *et al.* Coming to America: multiple origins of New World geckos. *J. Evol. Biol.* **24,** 231–244 (2011).
26. Mayr, G., Alvarenga, H. & Mourer-Chauviré, C. Out of Africa: fossils shed light on the origin of the hoatzin, an iconic Neotropic bird. *Naturwissenschaften* **98,** 961–966 (2011).

METHODS

No statistical methods were used to predetermine sample size.

The specimens were collected by wet screen-washing at the outcrop, followed by subsequent laboratory wet screening and separation into different size fractions (that is, >2 mm, <2 mm >1 mm, etc.). Treatment with 10% formic acid reduced the subsequent volume of matrix by ~85%. Further concentration of matrix using heavy liquid techniques and lithium metatungstate was attempted, but large quantities of manganese and iron compounds limited volume reduction. This problem was partly resolved by separating the heavier minerals before attempting to separate the fossils from the lighter minerals. *In situ* fracturing of fossil teeth[9] resulted in numerous partial specimens after processing.

The Santa Rosa primates are very small and rare. Whereas rodents and metatherians occur by the hundreds in the Santa Rosa local fauna, so far only four primates have been found. Discovery of the primates was delayed long after the initial description of the palaeofauna[8] because they only occur in the small size fraction of the matrix, the processing of which was not possible because of time constraints and the volume of matrix before discovery of the formic acid treatment. However, with the discovery of the first primate specimen in 2013, efforts to process the smaller-sized matrix were energized. Application of new techniques and funding made processing of smaller-sized material possible, and it is from this material that the primates have come.

The specimens described herein are permanently housed in the collections of the Área de Paleontología, Dirección de Geología Regional, INGEMMET, Lima, Peru. Casts are maintained in the Vertebrate Paleontology Department at the Natural History Museum of Los Angeles County.

LETTER

Phylogenetic structure and host abundance drive disease pressure in communities

Ingrid M. Parker[1,2]*, Megan Saunders[3], Megan Bontrager[1], Andrew P. Weitz[1], Rebecca Hendricks[3], Roger Magarey[4], Karl Suiter[4] & Gregory S. Gilbert[2,3]*

Pathogens play an important part in shaping the structure and dynamics of natural communities, because species are not affected by them equally[1,2]. A shared goal of ecology and epidemiology is to predict when a species is most vulnerable to disease. A leading hypothesis asserts that the impact of disease should increase with host abundance, producing a 'rare-species advantage'[3–5]. However, the impact of a pathogen may be decoupled from host abundance, because most pathogens infect more than one species, leading to pathogen spillover onto closely related species[6,7]. Here we show that the phylogenetic and ecological structure of the surrounding community can be important predictors of disease pressure. We found that the amount of tissue lost to disease increased with the relative abundance of a species across a grassland plant community, and that this rare-species advantage had an additional phylogenetic component: disease pressure was stronger on species with many close relatives. We used a global model of pathogen sharing as a function of relatedness between hosts, which provided a robust predictor of relative disease pressure at the local scale. In our grassland, the total amount of disease was most accurately explained not by the abundance of the focal host alone, but by the abundance of all species in the community weighted by their phylogenetic distance to the host. Furthermore, the model strongly predicted observed disease pressure for 44 novel host species we introduced experimentally to our study site, providing evidence for a mechanism to explain why phylogenetically rare species are more likely to become invasive when introduced[8,9]. Our results demonstrate how the phylogenetic and ecological structure of communities can have a key role in disease dynamics, with implications for the maintenance of biodiversity, biotic resistance against introduced weeds, and the success of managed plants in agriculture and forestry.

Plant pathogens can be important drivers of community diversity, structure and dynamics[1,2,10,11]. A basic premise of epidemiology is that pathogen transmission often increases with host density[12,13]. Density-dependent disease provides a mechanism for the maintenance of plant diversity in natural communities, in which locally uncommon species enjoy a rare-species advantage—based on lower enemy pressure—that mitigates the competitive impacts of dominant species[3–5]. Reports of density-dependent disease dynamics generally infer the potential effects on communities from studies of one or a few species[2], while community-level studies[1] are scarce but essential to evaluate whether such a rare-species advantage predicts patterns of disease across a community.

An ongoing debate concerns how community context influences disease, and particularly whether biodiversity suppresses infection and emerging diseases[14,15]. If increasing the number of species in a community reduces the density of competent hosts or the frequency of infected vectors, then biodiversity shows a suppressive 'dilution effect' on disease[16]. On the other hand, biodiversity may increase disease through a variety of mechanisms[15,17]. Pathogen spillover from one host species to

another means that community composition as well as species diversity can influence disease dynamics[18]. An unresolved question is whether these effects of community composition are idiosyncratic, or rather follow general rules that can be used to predict disease pressure.

The particular host species that are susceptible to a pathogen are a non-random subset of the local community, because a pathogen is more likely to be able to infect closely related species than evolutionarily distant ones[6,7,19]. This means that phylogenetic distance among species could be used to predict which local species are likely to be alternative hosts for the same pathogen[6]. By extension, species that are phylogenetically rare—that is, evolutionarily distant from species in their surrounding community—should suffer less disease pressure than other species (Fig. 1), as suggested for herbivory[20].

Working in a coastal grassland dominated by annual plants, we linked abundance to disease across all host species. We quantified the relative abundance of all 43 vascular plant species, which ranged over three orders of magnitude in per cent cover. Foliar pathogens are a ubiquitous aspect of plant life and can have a large impact on plant survival and productivity[21], including in these coastal grassland systems[22]. We measured disease pressure on every species as per cent diseased (necrotic or chlorotic) tissue. All species showed some necrotic or chlorotic disease symptoms; mean diseased leaf tissue varied from 1.4% to 37.5% across species. As predicted, disease pressure was significantly lower on rare species (Fig. 2a).

We then tested for an effect of community context on disease. We created a phylogeny for the community from a new supertree with stable ages for all vascular plants (Supplementary Information and Extended Data Fig. 1). We found that disease pressure on a species was inversely related not only to its numerical rarity but also to its phylogenetic rarity (Fig. 2b). We then constructed a tool to estimate how each member of a community could influence the epidemiology of pathogens on a given host species. We used a global database of associations among 210 host genera and 212 fungal pathogens to model the probability of sharing a pathogen as a function of the phylogenetic distance between two plant species (Extended Data Table 1 and Extended Data Fig. 2). This 'PhyloSusceptibility model' (pS model) was used to predict the probability of sharing a pathogen between all pairs of species in the grassland community. Because pathogen spillover should contribute to disease pressure, we tested whether the overall probability of sharing pathogens with neighbours was a good predictor of disease for each species.

Our accuracy in modelling the observed level of disease in each plant species improved when we included the expected influence of other plant species in the community. The most successful predictive model included the abundance of both the focal host and other species (Fig. 2d). We found that disease increased with the phylogenetically weighted abundance of other plants even in the absence of the effect of the focal host abundance (Fig. 2c). This demonstrates that pathogen

[1]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, California 95064, USA. [2]Smithsonian Tropical Research Institute, Apartado 2072, Balboa, República de Panamá. [3]Department of Environmental Studies, University of California Santa Cruz, Santa Cruz, California 95064, USA. [4]Center for Integrated Pest Management, North Carolina State University, Raleigh, North Carolina 27606, USA.
*These authors contributed equally to this work.

Figure 1 | Pathogen spillover depends on abundance and relatedness.
Schematic of the combined influence of phylogenetic distance and relative abundance of community members on pathogen spillover to a focal species. The focal species C is more likely to share pathogens with close relatives (D and E) than phylogenetically distant species (A, B and F), and hosts with greater abundance (indicated by number of leaves) are likely to produce more pathogen inoculum than locally rare species. Arrow thickness reflects the combined effects of phylogenetic distance and relative abundance on spillover from each alternative host in the community.

spillover from other hosts is important, and that phylogenetic structure and ecological structure both play a part in disease pressure across a host community (see Extended Data Fig. 3 for heuristic examples). The PhyloSusceptibility model and the associated stable supertree are tools that can be applied to any plant system to predict pathogen or pest pressure and to generate phylogenetic trees for communities that can be easily compared across studies.

The link between phylogeny and epidemiology implies that pathogens may promote the phylogenetic diversity of plant communities. Our grassland sites comprised assemblages with fewer close relatives than expected by chance from the local species pool (Extended Data Figs 4 and 5). This phylogenetic overdispersion at a local scale has also been observed by others[23], despite the expectation that phylogenetically constrained traits that influence plant distribution will lead to the clustering of related species[24]. The dominant explanation for phylogenetic overdispersion has been competition among similar species[23,24]. However, phylogenetically constrained sharing of pathogens could also provide a local advantage to distantly related species[25].

Furthermore, the role of community structure in disease pressure could help explain why some introduced plants become invasive while others do not. A central hypothesis in invasion ecology is that some species become aggressive weeds and pests because they leave their natural enemies behind in their home range[26,27]. At the same time, the great majority of introduced species do not become invasive, and disease pressure from resident species may contribute 'biotic resistance'[22,27]. Darwin's naturalization hypothesis suggests that those introduced species that are most successful are less closely related to residents[8], a pattern observed in California grasslands[9] although not universally supported[28]. One possible mechanism for this phenomenon is disease pressure originating from closely related resident species. The phylogenetic signal in host range predicts that pathogens that attack introduced species should come primarily from closely related residents, and by extension, introduced plants that are phylogenetically more isolated from the resident community are more likely to escape disease—a phylogenetic rare advantage.

We tested this prediction in our grassland with experimental introductions of novel species. We selected non-horticultural species not present in California from across the angiosperm phylogeny, representing varying degrees of phylogenetic relatedness to resident species (Extended Data Fig. 1). For each of 44 novel hosts, we predicted their susceptibility to local pathogens using the PhyloSusceptibility model and the species lists from the same ten grassland sites. We then grew plants in randomized arrays at the ten sites and quantified per cent symptomatic leaf tissue for each species. For ethical reasons, we removed all novel species before they reproduced.

Disease observed in the novel species was predicted remarkably well by the PhyloSusceptibility model integrated over the resident community (Fig. 3). The introduced species that escaped disease were all species with few close relatives, while those closely related to many residents always showed high levels of disease. In other words, spillover of pathogens from residents onto introduced hosts drove patterns of disease. The nonlinear response in Fig. 3 shows a saturating effect on disease that may reflect a functional limit, such as induced defences that reduce the spread
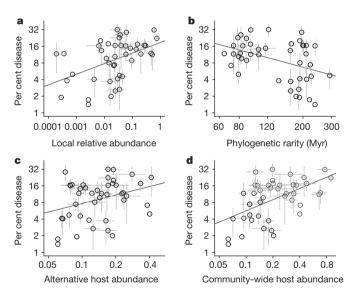
Figure 2 | The joint effects of abundance and phylogenetic signal on disease pressure. a, Per cent disease (necrosis and chlorosis) increased with the relative abundance of the focal species ($n = 43$, $r^2 = 0.22$, $P = 0.0017$). **b**, Per cent disease decreased with phylogenetic isolation from the rest of the community (unweighted by abundance), as measured by the 10th quantile of phylogenetic distance from the focal species ($n = 43$, $r^2 = 0.17$, $P = 0.0059$). Myr, million years. **c**, Pathogen spillover influenced disease on focal species; per cent disease was significantly predicted by the sum of the abundance of each other species (i) weighted by its phylogenetically determined probability of sharing a pathogen with the focal species, $p(S)_i$ ($n = 43$, $r^2 = 0.10$, $P = 0.037$). **d**, Disease was most strongly predicted by the focal species abundance combined with the $p(S)$-weighted abundance of all other species ($n = 43$, $r^2 = 0.28$, $P = 0.00023$). This community-wide, phylogenetically weighted host abundance (**d**; Akaike information criterion (AIC) = 26.6) was a better predictor of disease than the abundance of focal host alone (**a**; AIC = 30.5). Species means (± 1 standard error of the mean (s.e.m.)) across ten sites are shown; regressions are based on mean values. Points without error bars represent species measured in only one site.
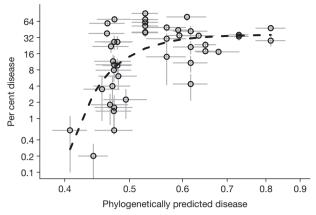
Figure 3 | Prediction of local disease from a global database. A global model of the probability of pathogen spillover as a function of unweighted phylogenetic distance from neighbouring species (PhyloSusceptibility, pS) significantly predicted per cent disease (necrosis and chlorosis) on experimentally introduced plants. The linear model is significant $(\log_{10}(\text{disease}) = 2.27 + 4.16 \times \log_{10}(pS))$, $n = 44$, $r^2 = 0.224$, $P = 0.0011$, residual standard error (RSE) = 0.57, AIC = 79.6), but a three-parameter nonlinear model provided a better fit (shown, $\log_{10}(\text{disease}) = 1.56 - 0.001$ $\exp(-18.76 \times \log_{10}(pS))$, RSE = 0.51, AIC = 71.5). Species means (± 1 s.e.m.) across ten sites are shown; regressions are based on mean values.

of pathogens into new host tissue, or a constraint on per cent disease imposed by the continuous production of new, uninfected leaves.

In a study at the continental scale, introduced plants that left more pathogens behind and accumulated fewer new pathogens were more likely to be invasive[26]. Our study is the first experimental demonstration, to our knowledge, of a local mechanism for variation in disease that could link Darwin's naturalization hypothesis to this continental pattern.

Crop introductions sometimes fail owing to overwhelming disease pressure from local pathogens[29]. The PhyloSusceptibility model could provide a useful tool for predicting when there is a high probability of elevated disease pressure on proposed horticultural introductions, novel crops or forestry species, biofuels plantations, or new intercropping combinations[6]. Similarly, we can quantify the relative vulnerability of local species of concern to pathogen spillover from these same introduced species. Models to predict the future distribution of plant invasions have been constrained by the difficulty of incorporating biotic drivers, including pathogens and parasites[30]. We show how fairly simple information on the species composition of resident plant communities could help predict both variation in the spread of an invasive species across a landscape and variation in the impact of that species.

Taken together, our results from wild communities and novel host introductions suggest that the structure of communities strongly influences host–pathogen interactions. The role of phylogenetic structure may help explain why loss of species sometimes increases and sometimes depresses disease, shedding new light on an important debate about the value of biodiversity. Finally, pathogens may have a key role in maintaining plant species diversity. Our results indicate that the rare-species advantage should also promote local phylogenetic diversity, leading to communities that capture a broader sampling of evolutionary history.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Bagchi, R. *et al.* Pathogens and insect herbivores drive rainforest plant diversity and composition. *Nature* **506,** 85–88 (2014).
2. Mordecai, E. A. Pathogen impacts on plant communities: unifying theory, concepts, and empirical work. *Ecol. Monogr.* **81,** 429–441 (2011).
3. Connell, J. in *Dynamics of Numbers in Populations* (eds den Boer, P. J. & Gradwell, G. R.) 298–312 (PUDOC, 1971).
4. Gillett, J. B. Pest pressure, an underestimated factor in evolution. *Syst. Ass. Publication* **4,** 37–46 (1962).
5. Janzen, D. Herbivores and the number of tree species in tropical forests. *Am. Nat.* **104,** 501–528 (1970).
6. Gilbert, G. S. & Webb, C. O. Phylogenetic signal in plant pathogen–host range. *Proc. Natl Acad. Sci. USA* **104,** 4979–4983 (2007).
7. Streicker, D. G. *et al.* Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science* **329,** 676–679 (2010).
8. Mack, R. N. in *Proceedings of the IX International Symposium on Biological Control of Weeds* (eds Moran, V. C. & Hoffman, J. H.) 39–46 (Univ. of Cape Town, 1996).
9. Strauss, S. Y., Webb, C. O. & Salamin, N. Exotic taxa less related to native species are more invasive. *Proc. Natl Acad. Sci. USA* **103,** 5841–5845 (2006).
10. Alexander, H. M. Disease in natural plant populations, communities, and ecosystems: insights into ecological and evolutionary processes. *Plant Dis.* **94,** 492–503 (2010).
11. Mangan, S. A. *et al.* Negative plant-soil feedback predicts tree-species relative abundance in a tropical forest. *Nature* **466,** 752–755 (2010).
12. Anderson, R. M. & May, R. M. Population biology of infectious diseases. Part 1. *Nature* **280,** 361–367 (1979).
13. Burdon, J. J. & Chilvers, G. A. Host density as a factor in plant disease ecology. *Annu. Rev. Phytopathol.* **20,** 143–166 (1982).
14. Wood, C. L. *et al.* Does biodiversity protect humans against infectious disease? *Ecology* **95,** 817–832 (2014).
15. Keesing, F. *et al.* Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* **468,** 647–652 (2010).
16. Johnson, P. T., Preston, D. L., Hoverman, J. T. & Richgels, K. L. Biodiversity decreases disease through predictable changes in host community competence. *Nature* **494,** 230–233 (2013).
17. Randolph, S. E. & Dobson, A. D. Pangloss revisited: a critique of the dilution effect and the biodiversity-buffers-disease paradigm. *Parasitology* **139,** 847–863 (2012).
18. Power, A. G. & Mitchell, C. E. Pathogen spillover in disease epidemics. *Am. Nat.* **164** (suppl. 5), S79–S89 (2004).
19. Wapshere, A. J. A strategy for evaluating the safety of organisms for biological weed control. *Ann. Appl. Biol.* **77,** 201–211 (1974).
20. Yguel, B. *et al.* Phytophagy on phylogenetically isolated trees: why hosts should escape their relatives. *Ecol. Lett.* **14,** 1117–1124 (2011).
21. Strange, R. N. & Scott, P. R. Plant disease: a threat to global food security. *Annu. Rev. Phytopathol.* **43,** 83–116 (2005).
22. Parker, I. M. & Gilbert, G. S. When there is no escape: the effects of natural enemies on native, invasive, and noninvasive plants. *Ecology* **88,** 1210–1224 (2007).
23. Swenson, N. G., Enquist, B. J., Thompson, J. & Zimmerman, J. K. The influence of spatial and size scale on phylogenetic relatedness in tropical forest communities. *Ecology* **88,** 1770–1780 (2007).
24. Webb, C. O., Ackerly, D. D., McPeek, M. A. & Donoghue, M. J. Phylogenies and community ecology. *Annu. Rev. Ecol. Syst.* **33,** 475–505 (2002).
25. Liu, X. B. *et al.* Experimental evidence for a phylogenetic Janzen-Connell effect in a subtropical forest. *Ecol. Lett.* **15,** 111–118 (2012).
26. Mitchell, C. E. & Power, A. G. Release of invasive plants from fungal and viral pathogens. *Nature* **421,** 625–627 (2003).
27. Elton, C. S. *The Ecology of Invasions by Animals and Plants* (Univ. of Chicago Press, 1958).
28. Duncan, R. P. & Williams, P. A. Darwin's naturalization hypothesis challenged. *Nature* **417,** 608–609 (2002).
29. Goodell, K., Parker, I. M. & Gilbert, G. S. in *Incorporating Science, Economics, and Sociology in Developing Sanitary and Phytosanitary Standards in International Trade* (ed. Caswell, J.) 87–117 (National Academy Press, 2000).
30. Wisz, M. S. *et al.* The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev. Camb. Philos. Soc.* **88,** 15–30 (2013).

Author Contributions I.M.P. and G.S.G. conceived and designed the study, managed the project, performed fieldwork and laboratory work, analysed the data, and wrote the paper. G.S.G. built the supertree R2G2_20140601 and the PhyloSusceptibility model. I.M.P. wrote the initial manuscript. M.S. helped design the study and performed laboratory work. M.B., A.P.W. and R.H. performed fieldwork and laboratory work. R.M. and K.S. contributed the database for the PhyloSusceptibility model.

METHODS

Habitat description. Our field sites were located in the extensive grassland associated with the campus of the University of California, Santa Cruz (UCSC) ($36°\,59'\,18.09''$ N, $122°\,3'\,31.29''$ W), on the central coast of California, USA[31]. This area experiences a Mediterranean climate, with a summer dry season and most precipitation coming in the form of winter rains. The area is strongly dominated by herbaceous plants, with only 2% of cover in shrubs. We documented 43 vascular plant species at this location; their abundances showed a characteristic lognormal distribution, with a few very abundant species and 69% of the species with <1% cover. The most common species were introduced Eurasian annual grasses that have dominated California grasslands for over 150 years[32]; *Avena barbata, Bromus diandrus* and *Brachypodium distachyon* together comprised over half of the plant cover.

Plant abundance. We quantified relative abundance of all 43 vascular plant species in the grassland. We located ten sites within the 30-hectare (ha) grassland area, navigating to randomly selected points 44–260 m apart. Each site was sampled in 2011 with a circular plot 20 m in diameter. Within each plot, we quantified per cent cover of all species in eight randomly placed, 50 cm × 100 cm rectangular quadrats. We chose per cent cover as our measure of host abundance rather than number of individual plants, because multiple, independent infections of leaves, combined with the modular growth of plants, point to abundance of plant tissue as the most appropriate measure of host density. From 11 April to 5 May 2011, we used a visual cover estimation method in which each quadrat was first divided into five 20 cm × 50 cm subsections. Two observers independently quantified cover of all species, and combinations of the same three people collected cover data for all ten sites, with high repeatability between observers (mean $r = 0.97$, $n = 81$ quadrats). We calculated post-hoc averages of the two independent cover measures for each site and species. Cover of rare species was estimated at the site scale (314 m^2) by measuring individuals or patches of all species found within the circular plot that were not found in more than two quadrats. Species were identified using the Jepson manual[33].

Disease pressure. Disease pressure was estimated for all species in the ten sites. In each site we laid four 10-m transects in the cardinal directions. At a point every 2 m along each transect, we collected the closest individual plant of all species within a 30-cm radius, thus the maximum sample size per species per site was 20. For any species with fewer than 5 individuals found on the transects, we searched the whole circular plot and collected haphazardly until we reached 5 individuals for that species. To ensure a measure of disease pressure across both young and old leaves, we collected the entire individual, or an entire branch for large perennials such as *Baccharis pilularis*. We then removed every living leaf from the plant, flattened, and glued them onto blue paper. We did not include senescent leaves. The same day, we scanned these pages of leaves at 300 dots per inch (dpi). We analysed scans using Assess: Image Analysis Software for Plant Disease Quantification[34] to estimate the percentage of leaf tissue affected by disease (necrosis and chlorosis).

From our previous extensive work with foliar diseases and herbivory in California grassland plants[22], we know that most necrosis and chlorosis is associated with ascomycete fungal pathogens (for example, *Stemphylium* spp., *Alternaria* spp., *Glomerella* spp., *Phomopsis* spp., *Cladosporium* spp., *Leptosphaerulina* spp.), plus rusts and viruses on the grasses. We carefully examined the scans of leaves at high magnification and excluded damage caused directly by herbivores (for example, necrotic damage from thrips or chewing damage). Although we did not measure the fitness effects of pathogens in this study, our previous work in coastal grasslands documented the effects of these same types of foliar pathogens using statistical modelling, experimental infection under controlled conditions, and fungicide experiments[22]. Because fungicides target different groups of pathogens and have differential effects across plant (and pathogen) species, we could not use a fungicide treatment to uniformly reduce disease pressure across all hosts in the community.

We chose as our measure of foliar disease the proportion of leaf tissue lost to necrosis and chlorosis, because it is directly comparable across host species, and because loss of photosynthetic tissue is a direct index of the impact of disease on host productivity. Our previous work demonstrated a quantitative link between foliar disease and fitness[22]. We captured the most common symptoms caused by above-ground pathogens in the grassland community but not the impacts of vascular wilts, systemic viruses, damping-off, or root rot. These types of diseases could not be incorporated into our assessment of foliar symptoms, but there are no a priori reasons to expect that the pathogens that cause those diseases would have fundamentally different phylogenetic responses than the foliar pathogens[35].

Spatial autocorrelation. We tested for spatial autocorrelation in species composition and overall disease pressure among the ten plots using Mantel tests in R. We calculated dissimilarity matrices among the plots using plant species abundance (quantitative Jaccard), species presence/absence (binary Jaccard), and mean overall disease (quantitative Jaccard). We calculated physical distance between each

pair of sites from UTM coordinates (interplot distances ranged from 44 m to 260 m). Mantel tests (based on Pearson's product-moment correlation and 999 permutations) showed no significant spatial autocorrelation for the quantitative Jaccard ($r = 0.140$, $P = 0.226$) or binary Jaccard ($r = 0.146$, $P = 0.17$) for species composition, or for overall disease pressure ($r = -0.067$, $P = 0.63$). In addition, there was no significant relationship between the Jaccard quantitative distances for species similarity and overall disease ($r = -0.054$, $P = 0.59$).

Experimental introductions. We created experimental plant introductions by setting out randomized arrays of novel hosts. We chose plant species that do not grow as native species in California and were not ever, to our knowledge, horticultural introductions in California. We selected herbaceous species from across the angiosperm phylogeny, and for locally important groups (Poaceae, Asteraceae, Fabaceae), we chose species from a number of different tribes. The species were selected to obtain a broad range of phylogenetic rarity in the sample. For each of our candidate experimental species, we used Phylocom (http://phylodiversity.net/phylocom/)[36] to calculate the phylogenetic distances to all resident species using the information on species composition collected in 2011. We then generated histograms to visualize the distributions of those distances. Our final suite of 73 novel species represented a range of phylogenetic similarity to the local community, characterized by the variety of shapes of these histograms. Avoiding horticultural species and plants bred for disease resistance, we ordered seeds from native plant nurseries at a range of locations: Prairie Moon Nursery in Winona County, Minnesota (http://www.prairiemoon.com), Native American Seed in Kimble County, Texas (http://www.seedsource.com), and Ohio Prairie Nursery in Portage County, Ohio (http://www.ohioprairienursery.com). Our final list of 44 species was the subset of these plants that could be germinated successfully and grown within our time frame.

Seeds were sown into flats on 13 January 2012, transplanted after germination into 3.8 cm × 14 cm Ray Leach cone-tainers with potting soil, and grown in the UCSC greenhouse. On 27 March 2012, randomized arrays of 44 species were set out in the field and watered from below using wicks in tubs of water. We used the same ten random locations from which relative abundance was estimated in 2011 and placed one completely randomized array at each site. On 19 April 2012, we quantified disease pressure on each plant, assessing per cent necrotic and chlorotic tissue according to the following categories: 0%, 1%, 5%, 10%, 20%, 50%, 70%, 95% and 100%. No leaves were necrotic at the start of the experiment. Because pathogen infection can increase rates of senescence[37], fully senescent leaves were included as 100% damaged in the results presented here. However, we obtained similar results when we excluded senescent leaves. All arrays were removed from the field before plants could set seed so that no biosafety risk was created by this project.

Estimating phylogenetic distances. We estimated pairwise phylogenetic distances among all plant species in the grassland community and experimentally introduced plants. To make our approach as useful and generalizable as possible, we hand constructed a phylogenetic supertree, R2G2_20140601, with stable ages for all vascular plant families (see Supplementary Information for details of tree construction and sources, as well as a dated Newick file of the full tree). We then used Phylomatic v.4.2 and the Phylocom bladj function (http://phylodiversity.net/phylocom/)[36] to create a dated phylogenetic tree for all taxa in the grassland community and experimentally introduced plants, based on the R2G2_20140601 tree (Extended Data Fig. 1). We calculated pairwise phylogenetic distances among all pairs of species with the cophenetic function in the Picante Package of the R Statistical Framework. Phylogenetic distances are in millions of years (Myr) of independent evolution, which is equivalent to twice the time to most recent common ancestor. We estimated phylogenetic rarity as the 10th quantile distance to other community members in each site, averaged over the ten sites. The 10th quantile was chosen because other measures such as nearest neighbour distance, mean distance, and 25th quantile did not provide as much variability among taxa within which to distinguish patterns. There was no statistical correlation between phylogenetic rarity and local abundance (Spearman $\rho = 0.0014$, $P = 0.93$), which simplifies interpretation of the data.

Evaluating phylogenetic dispersion in plant communities. We compared the phylogenetic dispersion of species within the grassland plots to that expected if species were drawn randomly from a regional metacommunity pool (predicting phylogenetic clustering from habitat filtering) or to the pool of species in the overall grassland community (predicting overdispersion through negative species interactions). For each of the ten sites ($n = 15$–32 species per site), we randomly sampled the same species richness from either the 529 vascular plant species found on the surrounding region of the 810-ha UCSC Campus (including grassland, forest and chaparral habitats[31]) or from the pool of 43 species found across the ten grassland sites. We then calculated the phylogenetic distances (as described earlier) among all pairs of species within a plot (observed) and within each of 9,999 random draws. For each of the 10,000 samples, we calculated a range of quantiles of the pairwise phylogenetic distances from 0% (nearest taxon distance) to 50%

(median distance). The number of species pairs in a plot is $[(n \times n) - n]/2$ where n is the species richness, so a plot with 15 species has 105 phylogenetic distances and a plot with 32 species has 496. We then calculated the difference between each of the phylogenetic distance quantiles from the observed species assemblage to that from each random draw. This difference will be negative if the observed distances are shorter than expected (phylogenetic clustering) and positive if the observed distances are phylogenetically overdispersed (Extended Data Fig. 4). We then recorded the median difference for each of the quantiles for each of the ten sites. We tested whether the phylogenetic distances were greater or less than expected at random using a one-sample t-test (degrees of freedom (df) $= 9$) against the expected difference of 0 Myr (Extended Data Fig. 5). This approach has an advantage over the more commonly used mean nearest taxon distance or mean pairwise distance metrics[24] because it can uncover shifts in phylogenetic dispersion at many levels.

PhyloSusceptibility estimation of host sharing probabilities. From an independent global data set of associations among 210 plant genera and 212 fungal pathogens, we parameterized a PhyloSusceptibility (pS) model of the probability of sharing a pathogen as a function of phylogenetic distance. Previous work established a broad pattern of phylogenetic signal in host ranges of plant pests and pathogens[6,35]. We re-analysed the same host records for 1,670 pest species (including 212 fungal pathogens) on 210 angiosperm genera from the US Department of Agriculture Global Pest and Disease Database as described previously[35], but using the updated, finer resolution R2G2_20140601 tree with stable dates. We used the same logistic-regression analytical approach as described previously[35], with two exceptions. First, we included all pest species, even those reported from only one host (the 2012 study excluded these specialists, because it focused on polyphagous species). Second, we included the case of zero phylogenetic distance (distance from a known host species to itself). Both of these changes, as well as the finer-scale resolution of the phylogenetic tree, make the estimates of phylogenetic signal more generally applicable and steeper—that is, the probability of a pest being found on two hosts declines more rapidly with phylogenetic distance between the hosts than the estimates from ref. 35. Logistic regressions provide estimates of the probability that any two species of plants would share a pest given the phylogenetic distance between them, following the form $\mathrm{logit}(S) = \beta_0 + \beta_1 \times \log_{10}(\mathrm{PD} + 1)$, where S is whether the target host was susceptible to a pest known from the source host, and PD is the phylogenetic distance (time of independent evolution in Myr) between the source and target host. The probability that the target host is susceptible is then $\mathrm{p}(S) = \exp[\mathrm{logit}(S)]/[1 + \exp(\mathrm{logit}(S))]$. The logistic regression coefficients and confidence intervals for each of nine pest groups is provided in Extended Data Table 1; for fungal pathogens, $\beta_0 = 3.36$, $\beta_1 = -2.86$. The data matrix used for this analysis is available in supplementary data from ref. 35.

For the experimental arrays, we then used this pairwise probability to predict the overall probability that each novel species would share a pathogen with any species in the resident community. A focal species has a probability $\mathrm{p}(S)_i$ of sharing

a pest with each other species i; the probability that the focal species shares a pest with any resident species can then be calculated as the complement of the product of the probabilities of not sharing a pathogen with each other species $(1 - \mathrm{p}(S)_i)$ using the equation:

$$1 - \prod_{i=1}^{n-1}(1 - \mathrm{p}(S)_i)$$

for a community with n host species total. We call this the PhyloSusceptibility (pS) model.

Prediction of disease. For the wild plant community, we tested the predictive power of conspecific abundance, phylogenetic rarity and community structure (the abundance of all other species weighted by their phylogenetic distance) in explaining disease. We used linear regression to test the response of ($\log_{10}$) proportion diseased leaf tissue to ($\log_{10}$) relative abundance of the focal host, and to the 10th quantile phylogenetic distance to the other plant species (without weighting for relative abundance). We also calculated 'alternative host abundance' by multiplying the abundance of each other plant species i by $\mathrm{p}(S)_i$, the probability of sharing pathogens between that species and the focal species (pathogen spillover), then taking the sum over all other species. We added the focal species abundance to alternative host abundance to generate 'community-wide host abundance'. Disease and abundance measures were determined for each of the ten sites and averaged before analysis. We used linear regression to test the ability of ($\log_{10}$) alternative host abundance and ($\log_{10}$) community-wide host abundance to predict ($\log_{10}$) proportion disease.

For the experimental introduction experiment, we fit a nonlinear regression model to quantify the power of the pS model for predicting ($\log_{10}$) proportion diseased tissue on the novel species. Including local abundance of the resident species did not improve the model fit (data not shown). Log transformations were used to improve normality and heteroscedasticity. All analyses were done in R v.3.0.2.

31. Haff, T. M., Brown, M. T. & Tyler, W. B. *The Natural History of the UC Santa Cruz Campus* 2nd edn (Environmental Studies, UC Santa Cruz, 2008).
32. Mack, R. N. in *Biological Invasions: A Global Perspective* (ed. Drake, J. A.) 155–179 (John Wiley & Sons, 1989).
33. Baldwin, B. G. & Goldman, D. H. *The Jepson Manual: Vascular Plants of California* 2nd edn (Univ. of California Press, 2012).
34. Assess 2.0. *Image Analysis Software for Plant Disease Quantification v. 2.0* (APS Press, 2008).
35. Gilbert, G. S., Magarey, R., Suiter, K. & Webb, C. O. Evolutionary tools for phytosanitary risk analysis: phylogenetic signal as a predictor of host range of plant pests and pathogens. *Evol. Appl.* **5,** 869–878 (2012).
36. Webb, C. O. & Donoghue, M. J. Phylomatic: tree assembly for applied phylogenetics. *Mol. Ecol. Notes* **5,** 181–183 (2005).
37. Gilbert, G. S. & Parker, I. M. Rapid evolution in a plant–pathogen interaction and the consequences for introduced host species. *Evol. Appl.* **3,** 144–156 (2010).

LETTER

Agrochemical control of plant water use using engineered abscisic acid receptors

Sang-Youl Park[1,2]*, Francis C. Peterson[3]*, Assaf Mosquna[1,2]†*, Jin Yao[1,2], Brian F. Volkman[3] & Sean R. Cutler[1,2]

Rising temperatures and lessening fresh water supplies are threatening agricultural productivity and have motivated efforts to improve plant water use and drought tolerance. During water deficit, plants produce elevated levels of abscisic acid (ABA), which improves water consumption and stress tolerance by controlling guard cell aperture and other protective responses[1,2]. One attractive strategy for controlling water use is to develop compounds that activate ABA receptors, but agonists approved for use have yet to be developed. In principle, an engineered ABA receptor that can be activated by an existing agrochemical could achieve this goal. Here we describe a variant of the ABA receptor PYRABACTIN RESISTANCE 1 (PYR1) that possesses nanomolar sensitivity to the agrochemical mandipropamid and demonstrate its efficacy for controlling ABA responses and drought tolerance in transgenic plants. Furthermore, crystallographic studies provide a mechanistic basis for its activity and demonstrate the relative ease with which the PYR1 ligand-binding pocket can be altered to accommodate new ligands. Thus, we have successfully repurposed an agrochemical for a new application using receptor engineering. We anticipate that this strategy will be applied to other plant receptors and represents a new avenue for crop improvement.

The phytohormone ABA (Fig. 1a) has an essential role in regulating plant water use and drought tolerance. A land-plant-specific signalling network composed of receptors, phosphatases and kinases mediates ABA responses[1]. ABA receptors control the activity of a subfamily of three SNF1-related protein kinases (SnRK2 kinases) in response to environmental stress. These SnRK2 kinases autoactivate by *cis*- and *trans*-autophosphorylation on their activation loops[3,4], but are continuously inactivated by type 2C protein phosphatases (clade A PP2Cs), which results in low basal kinase activity. When ABA levels rise during stress, the phytohormone binds to soluble ABA receptors and stabilizes their activated conformations, enabling them to bind to and inhibit PP2Cs[5–11]. This in turn allows accumulation of activated SnRK2 kinases, whose direct targets include SLOW ANION CHANNEL 1, an anion channel that controls guard cell aperture, and ABA RESPONSE-ELEMENT-BINDING FACTORS, b-ZIP transcription factors that mediate ABA-regulated gene expression[12]. Thus, ABA controls water use and stress physiology by receptor-mediated inhibition of PP2C activity and resultant SnRK2 kinase activation.

Plants regulate their transpiration rates by modifying stomatal aperture, and consequently ABA receptors have emerged as attractive targets for water use optimization; however, ABA agonists approved for this use have yet to be developed. We reasoned that agrochemical control of plant water use could be accomplished in transgenic plants that express an engineered ABA receptor that responds to an existing agrochemical, a strategy based on orthogonal ligand–receptor systems, which have enabled selective chemical control of diverse targets[13,14]. To identify ligands suitable for our strategy, we constructed a collection of PYR1 mutants with saturating mutations in ligand-contacting residues and screened it to identify activating agrochemical ligands. Once identified, the sensitivities

of specific ligand–receptor pairs were improved through targeted mutagenesis and functional selection. This scheme was facilitated by a previously constructed set of PYR1 mutants that contains all possible 475 single amino acid substitutions in the 25 residues that line the ABA binding pocket[15]. Since overexpression of wild-type receptors has negative yield consequences[16], we inactivated the intrinsic ABA responsiveness of each of the 475 mutants by introducing an arginine at position K59, a highly conserved residue that forms a salt bridge with the carboxylate of ABA in wild-type receptors[7–11]. Each member of the mutant collection was individually tested for responsiveness to a panel of 15 commonly used non-herbicidal agrochemicals at high concentrations (100 μM) using a yeast two-hybrid-based assay that measures agonist-induced binding of receptor to PP2C[6,17]. This screening effort, which involved 7,125 mutant receptor–ligand response assays, identified receptors weakly responsive to 4 of the 15 compounds tested (Extended Data Fig. 1). This high hit rate is probably a consequence, in part, of the intrinsically low basal activity of PYR1, which facilitated the identification of weak responders.

We next attempted to optimize response sensitivities using targeted mutagenesis and functional selections. This worked most successfully
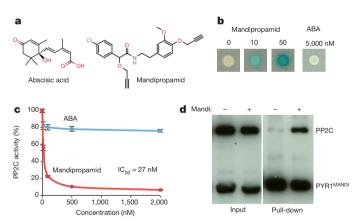

Figure 1 | PYR1^MANDI possesses nanomolar sensitivity to mandipropamid and functions *in vitro* and *in vivo*. a, Structures of ABA and mandipropamid. **b**, PYR1^MANDI binds to the PP2C HAB1 in response to mandipropamid, but not ABA, as measured using a yeast two-hybrid assay. **c**, PYR1^MANDI inhibits PP2C phosphatase activity in response to mandipropamid (red line), but not ABA (blue line). Shown are data using ABI1; ABI2 and HAB1 were also tested (IC$_{50}$ = 76 and 32 nM respectively). **d**, 6×His–GFP–PYR1^MANDI and GFP–HAB1 were co-expressed in *N. benthamiana* leaves using *Agrobacterium tumefaciens*. Plants were treated with mock (−) or 50 μM mandipropamid (Mandi; +) solutions. Twenty hours later PYR1^MANDI was affinity purified from treated leaves. Input extracts and affinity-purified proteins were separated by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and detected using an anti-GFP antibody.

[1]Center for Plant Cell Biology and Department of Botany and Plant Sciences, University of California, Riverside, California 92521, USA. [2]Institute for Integrative Genome Biology, Riverside, California 92521, USA. [3]Department of Biochemistry, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. †Present address: Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, Faculty of Agriculture, Hebrew University of Jerusalem, Rehovot 7610001, Israel.
*These authors contributed equally to this work.

with receptors responsive to mandipropamid (Fig. 1a), a mandelamide compound used to control oomycete (blight) pathogens that is sold under the trade name Revus. Mutations in four separate residues (in combination with K59R) were initially observed to confer mandipropamid responsiveness (Extended Data Figs 1 and 2). We constructed combinations of a subset of these mutations and identified a triple mutant (K59R, S122G, F108A) with low micromolar responsiveness in yeast-based assays (Extended Data Fig. 2). This mutant was next subjected to saturation mutagenesis at 22 pocket-lining residues and separately mutagenized by DNA shuffling, which together identified mutations in five residues that individually enhance sensitivity of the triple mutant (Extended Data Fig. 2). These enhancing mutations were assembled combinatorially using multiple site-directed mutagenesis and the mutagenized receptors were characterized directly. These efforts yielded a hextuple mutant, PYR1^{MANDI} (PYR1(Y58H/K59R/V81I/F108A/S122G/F159L)) that possesses nanomolar mandipropamid sensitivity in *in vitro* PP2C inhibition assays (half-maximum inhibitory concentration (IC_{50}) = 27 nM; Fig. 1b, c). To rule out potential artefacts caused by our reliance on yeast-based receptor activation assays and selections, we examined PYR1^{MANDI} activity in *Nicotiana benthamiana*. PYR1^{MANDI}, but not PYR1, binds to the PP2C HYPERSENSITIVE TO ABA1 (HAB1) in response to mandipropamid when both proteins are co-expressed in *N. benthamiana*, demonstrating that PYR1^{MANDI} functions in a plant cell environment (Fig. 1d). PYR1^{MANDI} is not activated by ABA *in vitro*, nor does mandipropamid substantially activate wild-type PYR1 or ten other *Arabidopsis* ABA receptors tested (Extended Data Table 1). Thus, PYR1^{MANDI} is selectively activated by mandipropamid.

To understand the molecular basis for the engineered agrochemical sensitivity we solved the X-ray crystal structure of a quadruple mutant receptor, PYR1(K59R/V81I/F108A/F159L), in complex with mandipropamid and HAB1 at 2.25 Å resolution (data collection structural statistics are shown in Extended Data Table 2). This quadruple mutant contains four of the six mutations present in PYR1^{MANDI} and possesses nanomolar sensitivity to mandipropamid (IC_{50} = 50 nM; Extended Data Fig. 3A); it was used because it formed higher-quality crystals than could be obtained with the hextuple mutant PYR1^{MANDI}. The mutant

receptor binds the (*S*)-stereoisomer of mandipropamid, which adopts a U-shaped orientation (Extended Data Fig. 3B, C) reminiscent of sulfonamide ABA receptor agonists[17–21] and induces a closed-gate receptor conformation (Extended Data Fig. 3D) that is nearly indistinguishable from previously determined PYR1–agonist–PP2C complexes (Cα root mean squared deviation (r.m.s.d.) of ~0.45 Å).

The structure obtained reveals how the ABA-binding pocket of PYR1 was transformed to bind mandipropamid. The most conspicuous change in the mutant receptor is an increase in the volume of the ligand-binding pocket created by two mutations, F108A and F159L, that enable mandipropamid's lengthy propargyl substituents to fit in the binding pocket (Fig. 2). F108A and F159L also enable hydrophobic contacts between mandipropamid and S109 and between mandipropamid and the G392 of HAB1, neither of which makes ABA contacts in wild-type structures (Extended Data Fig. 4). In addition, the K59R mutation enables a hydrogen bond between the Nε of R59 and the amide carbonyl of mandipropamid, mimicking the direct contact between K59 and ABA's carboxylate in wild-type PYR1 (Extended Data Fig. 4). A second direct hydrogen bond occurs between E94 and mandipropamid's amide carbonyl oxygen. Binding is also stabilized by extensive hydrophobic contacts to residues that would normally contact ABA in wild-type PYR1 and two separate water-mediated hydrogen-bond networks that interact with mandipropamid's amide carbonyl or its 3- and 4-alkoxy phenethyl substituents and the Trp lock (Extended Data Fig. 4). The contributions that the Y58H, V81I and S122G mutations present in PYR1^{MANDI} make to ligand binding are less obvious, as they occur in residues too distant to make direct mandipropamid contacts. Collectively, these structural observations show that a relatively small number of mutations are sufficient to reshape PYR1's ligand-binding pocket so that it can be activated by an unnatural ligand.

An unusually large family of receptors binds ABA in land plants. The restricted activation of PYR1 and closely related receptors by the sulfonamide agonist quinabactin is sufficient to induce a full ABA response, indicating that simultaneous activation of all 14 ABA receptors is not necessary for synthetic pathway activation[20,21]. It is likely, although untested, that activating PYR1 should be sufficient to elicit global pathway
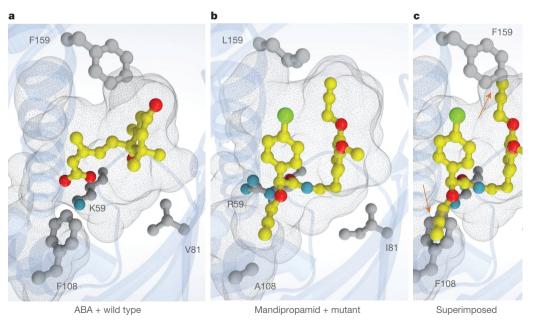

Figure 2 | Crystal structure of a mandipropamid responsive receptor— F108A and F159L prevent steric clash. a, Previously published coordinates (Protein Data Bank accession 3QN1) were used to represent the inner surface of wild-type PYR1, which is shown as a mesh enclosing ABA (shown in yellow); the side chains altered in the mandipropamid-responsive mutant are shown in grey. **b**, X-ray coordinates for a PYR1(K59R/V81I/F108A/F159L)– mandipropamid–HAB1 complex were obtained experimentally and used to

represent the ligand-binding pocket. Mandipropamid (yellow) is sold as mixed stereoisomers, but the mutant receptor selectively binds the (*S*)-isomer. **c**, Superimposition of mandipropamid (yellow) onto the wild-type receptor shows that the wild-type receptor disfavours binding due to steric cash with F108 and F159 (red arrows). Structures were rendered in Cinema4D using ePMV[23]. The inner surface meshes shown were exported from PyMol; the latch loop has been omitted for clarity.
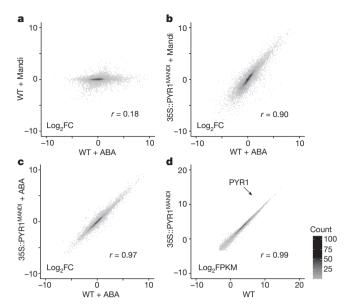
Figure 3 | Mandipropamid induces an ABA-like transcriptional response selectively in the PYR1^MANDI genotype. a–d, The wild-type (WT) and transgenic PYR1^MANDI genotypes were treated with 50 μM ABA, 2 μM mandipropamid (Mandi) or mock solutions for 8 h in biological triplicate. RNA was then isolated and used for RNA-seq experiments. **a**, Mandipropamid does not induce and ABA-like transcriptional response in the wild type genotype. **b**, Mandipropamid induces an ABA-like effect in the PYR1^MANDI genotype. **c**, The PYR1^MANDI genotype responds normally to ABA. **d**, The PYR1^MANDI transgene does not substantially alter basal transcript levels in the absence of mandipropamid treatment. **a–c**, Hexbin plots of $\log_2$-transformed fold change (FC) (chemical/mock) values for ~21,000 transcripts with fragments per kilobase of exon per million fragments mapped (FPKM) abundance values above 0.1 in all experimental samples; **d**, $\log_2$ FPKM values are plotted.

activation in adult plants. To test this, we made and characterized transgenic *Arabidopsis* plants that express PYR1^MANDI under the control of the constitutive viral 35S promoter. Seed germination is classically used to assess ABA effects, owing to ABA's role in inhibiting germination under environmentally unfavourable conditions. Mandipropamid inhibits seed germination in two independent 35S::PYR1^MANDI transgenic, but not wild-type or 35S::GFP–PYR1, strains. As expected, the mandipropamid sensitivity of the transgenic strains constructed correlates with PYR1^MANDI protein abundance (Extended Data Fig. 5A–C). Mandipropamid also inhibits primary root growth in the 35S::PYR1^MANDI strains, but not wild-type or 35S::GFP–PYR1 genotypes (Extended Data Fig. 5D). These data show that activating PYR1^MANDI is sufficient to control seed and root ABA responses.

To characterize the ABA response of 35S::PYR1^MANDI transgenic lines more closely, we used RNA sequencing (RNA-seq) experiments to compare the transcriptional responses induced by mandipropamid and ABA treatments in both the wild-type and 35S::PYR1^MANDI strains. As shown in Fig. 3a, mandipropamid does not induce a substantial ABA response in wild-type non-transgenic plants (Pearson's correlation coefficient $r = 0.17$); however, it does induce a global ABA-like response in the 35S::PYR1^MANDI line ($r = 0.90$; Fig. 3b). Additionally, the transcriptional responses of the wild-type and 35S::PYR1^MANDI genotypes to ABA treatments are highly correlated ($r = 0.97$), indicating that the 35S::PYR1^MANDI transgene does not interfere substantially with the endogenous ABA response mediated by wild-type receptors (Fig. 3c). The basal transcript levels of untreated wild-type and 35S::PYR1^MANDI genotypes are also highly correlated ($r = 0.99$; Fig. 3d), which indicates that the transgene has minimal effects on transcript abundances in the absence of mandipropamid treatment. Consistent with this, we observe negligible differences between the fresh weights or flowering times of the wild-type and two 35S::PYR1^MANDI genotypes grown in the absence
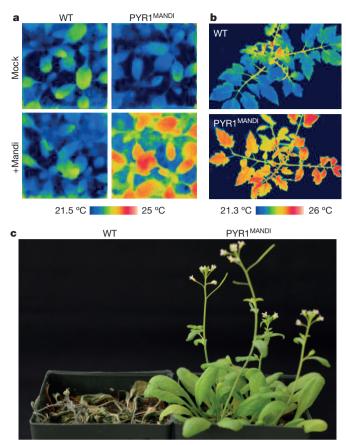
Figure 4 | Agrochemical control of transpiration and drought tolerance in the PYR1^MANDI genotype. a–c, A ~2 °C increase in leaf temperature is selectively observed in response to mandipropamid in the PYR1^MANDI genotype in *Arabidopsis* and tomato. WT, wild type. **a**, Three-week-old *Arabidopsis* seedlings were treated with 1 μM mandipropamid (Mandi) or a mock solution and imaged by thermography 24 h after application. Leaf warming is a consequence of reduced transpiration. **b**, Transgenic 35S::PYR1^MANDI tomato plants were grown alongside wild-type controls and treated with 25 μM mandipropamid and thermographed 24 h after application. **c**, Induction of drought tolerance by mandipropamid treatments of the PYR1^MANDI genotype. Three-week-old wild-type or PYR1^MANDI genotype plants were treated with mandipropamid twice over the course of an 11-day water deprivation period. Photographs were taken 24 h after re-watering. Drought survival experiments were conducted on three separate occasions with each experiment conducted using a minimum of three biological replicates. The data shown for each figure panel are subsets of larger experiments shown completely in Extended Data Figs 6–8.

of mandipropamid treatment (Extended Data Fig. 5E, F). Thus, mandipropamid induces a genome-wide ABA-like transcriptional response selectively in transgenic *Arabidopsis* plants expressing PYR1^MANDI and the expression of PYR1^MANDI is not associated with substantial changes in basal transcript levels or background ABA responsiveness.

A critical physiological role of ABA is to control guard cell aperture and transpiration rates. This can be measured indirectly through leaf temperature, which increases when guard cells close owing to decreased evaporative cooling. After treatment with mandipropamid, transgenic *Arabidopsis* plants expressing PYR1^MANDI show elevated leaf temperatures (Fig. 4a and Extended Data Fig. 6), indicating that PYR1^MANDI can function in guard cells. The effects of mandipropamid persist ~2 days longer than those of ABA for wild-type plants (Extended Data Fig. 6), which could be due to multiple reasons, including differences in metabolism between ABA and mandipropamid. To establish whether PYR1^MANDI is active in other species, we constructed transgenic 35S::PYR1^MANDI tomato plants and observed that they too show a similar increase in leaf temperature in response to mandipropamid treatments (Fig. 4b and

Extended Data Fig. 7), indicating that PYR1^{MANDI} can control transpiration rates in two divergent eudicotyledonous species. Furthermore, like the effects of ABA on wild-type plants, the action of mandipropamid on the 35S::PYR1^{MANDI} genotypes is sufficient to improve *Arabidopsis* survival after water deprivation (one measure of drought tolerance). When we subjected the wild-type, 35S::GFP–PYR1 and 35S::PYR1^{MANDI} genotypes to a water deprivation regime, mandipropamid treatments selectively improved survival in two independent 35S::PYR1^{MANDI} transgenic lines in three separate experiments (Fig. 4c and Extended Data Fig. 8), as expected based on the broad activation of ABA responses we have demonstrated with the PYR1^{MANDI}/mandipropamid system.

Our data demonstrate selective agrochemical control of ABA signalling in *Arabidopsis* using an engineered receptor and illustrate the power of synthetic biological approaches for manipulating plant physiology. PYR1^{MANDI} is a new tool that can be used to control water use and to probe the ABA response pathway. Given the relatively simple structural basis underlying PYR1^{MANDI} function, we anticipate that it should be possible to modify other ABA receptors so that individual family members can be selectively activated, which will facilitate functional analyses at the whole-plant and cell-type-specific levels. Our work also has biotechnological implications. Although the genetic manipulation of ABA responses has been validated in the field as a strategy for improving drought tolerance in canola[22], the broad use of the ABA pathway for manipulating drought tolerance is a relatively new idea that requires further validation. Moreover, the specific orthogonal control strategy outlined here will require testing in crops before it is suitable for use in the field. Nonetheless, our work demonstrates that it is possible to repurpose an existing agrochemical using receptor engineering. This strategy can be broadly applied to other plant receptors and agrochemicals and therefore opens new avenues for crop improvement.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Cutler, S. R., Rodriguez, P. L., Finkelstein, R. R. & Abrams, S. R. Abscisic acid: emergence of a core signaling network. *Annu. Rev. Plant Biol.* **61,** 651–679 (2010).
2. Kim, T. H., Böhmer, M., Hu, H., Nishimura, N. & Schroeder, J. I. Guard cell signal transduction network: advances in understanding abscisic acid, CO_2, and Ca^{2+} signaling. *Annu. Rev. Plant Biol.* **61,** 561–591 (2010).
3. Soon, F. F. *et al.* Molecular mimicry regulates ABA signaling by SnRK2 kinases and PP2C phosphatases. *Science* **335,** 85–88 (2012).
4. Ng, L.-M. *et al.* Structural basis for basal activity and autoactivation of abscisic acid (ABA) signaling SnRK2 kinases. *Proc. Natl Acad. Sci. USA* **108,** 21259–21264 (2011).
5. Ma, Y. *et al.* Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science* **324,** 1064–1068 (2009).
6. Park, S.-Y. *et al.* Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* **324,** 1068–1071 (2009).
7. Yin, P. *et al.* Structural insights into the mechanism of abscisic acid signaling by PYL proteins. *Nature Struct. Mol. Biol.* **16,** 1230–1236 (2009).
8. Nishimura, N. *et al.* Structural mechanism of abscisic acid binding and signaling by dimeric PYR1. *Science* **326,** 1373 (2009).
9. Santiago, J. *et al.* The abscisic acid receptor PYR1 in complex with abscisic acid. *Nature* **462,** 665–668 (2009).
10. Melcher, K. *et al.* A gate–latch–lock mechanism for hormone signalling by abscisic acid receptors. *Nature* **462,** 602–608 (2009).
11. Miyazono, K. *et al.* Structural basis of abscisic acid signalling. *Nature* **462,** 609–614 (2009).
12. Miyakawa, T., Fujita, Y., Yamaguchi-Shinozaki, K. & Tanokura, M. Structure and function of abscisic acid receptors. *Trends Plant Sci.* **18,** 259–266 (2013).
13. Bishop, A. *et al.* Unnatural ligands for engineered proteins: new tools for chemical genetics. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 577–606 (2000).
14. Belshaw, P. J., Schoepfer, J. G., Liu, K.-Q., Morrison, K. L. & Schreiber, S. L. Rational design of orthogonal receptor–ligand combinations. *Angew. Chem. Int. Edn Engl.* **34,** 2129–2132 (1995).
15. Mosquna, A. *et al.* Potent and selective activation of abscisic acid receptors *in vivo* by mutational stabilization of their agonist-bound conformation. *Proc. Natl Acad. Sci. USA* **108,** 20838–20843 (2011).
16. Kim, H. *et al.* Overexpression of PYL5 in rice enhances drought tolerance, inhibits growth, and modulates gene expression. *J. Exp. Bot.* **65,** 453–464 (2014).
17. Peterson, F. C. *et al.* Structural basis for selective activation of ABA receptors. *Nature Struct. Mol. Biol.* **17,** 1109–1113 (2010).
18. Melcher, K. *et al.* Identification and mechanism of ABA receptor antagonism. *Nature Struct. Mol. Biol.* **17,** 1102–1108 (2010).
19. Yuan, X. *et al.* Single amino acid alteration between valine and isoleucine determines the distinct pyrabactin selectivity by PYL1 and PYL2. *J. Biol. Chem.* **285,** 28953–28958 (2010).
20. Okamoto, M. *et al.* Activation of dimeric ABA receptors elicits guard cell closure, ABA-regulated gene expression, and drought tolerance. *Proc. Natl Acad. Sci. USA* **110,** 12132–12137 (2013).
21. Cao, M. *et al.* An ABA-mimicking ligand that reduces water loss and promotes drought resistance in plants. *Cell Res.* **23,** 1043–1054 (2013).
22. Wang, Y. *et al.* Molecular tailoring of farnesylation for plant drought tolerance and yield protection. *Plant J.* **43,** 413–424 (2005).
23. Johnson, G. T., Autin, L., Goodsell, D. S., Sanner, M. F. & Olson, A. J. ePMV embeds molecular modeling into professional animation software environments. *Structure* **19,** 293–303 (2011).

Author Contributions S.-Y.P. and A.M. conducted protein mutagenesis experiments. S.-Y.P. conducted and J.Y. analysed the RNA-seq experiments. F.C.P. conducted the protein crystallography experiments. S.-Y.P. constructed and analysed transgenic plants. B.F.V. and S.R.C. designed and supervised experiments collaboratively with all co-authors. S.R.C. conceived the project and wrote the manuscript with input from all co-authors.

Author Information The X-ray crystallographic coordinates and structure factor files for the engineered PYR1 mandipropamid receptor in complex with mandipropamid and HAB1 have been deposited in the Protein Data Bank under accession number 4WVO. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R.C. (sean.cutler@ucr.edu).

METHODS

Construction of the K59R site saturation mutagenized 'pocket' library. Site-saturation mutagenesis involves directly constructing all 19 possible single amino acid substitution mutations at target residues of interest and enables systematic coverage of substitutions at sites of interest[24]. We constructed a library of site-saturated mutants at 25 pocket-lining residues in the ABA non-responsive PYR1(K59R) backbone. In preliminary experiments screening error-prone PCR-mutagenized wild-type PYR1 templates for mutants that would respond to agrochemicals, we isolated the K59R mutation in screens against structurally dissimilar agrochemicals. It is possible that, in addition to eliminating PYR1's sensitivity to ABA, that the K59R mutation may sensitize PYR1 to non-specific chemical activation.

A set of site-saturated mutations was previously constructed in a wild-type PYR1 backbone as part of a larger effort focused on engineering constitutively active ABA variants[15]. The PYR1 template mutagenized in those experiments was a pBD-PYR1 plasmid that encodes a GAL4–DNA binding domain (BD) fusion to PYR1; this plasmid can be directly used for assaying receptor–PP2C interactions in an appropriate yeast strain co-transformed with pACT-PP2C, which express a GAL4 activation domain fusion (ACT) to a PP2C of interest[6]. We used pACT-HAB1 in the experiments shown.

We incorporated the K59R mutation into each of the original PYR1 wild-type backbone mutants using PCR-based mutagenesis, which yielded a collection of 475 PYR1(K59R) mutants in the following sites: P55, F61, I62, V81, V83, L87, P88, A89, S92, E94, E141, F108, I110, H115, R116, L117, Y120, S122, M158, F159, A160, T162, V163, V164 and N167. This was accomplished in two ways. Plasmids containing mutations in twenty-two of the sites targeted (all except P55, F61 and I62) were mutagenized using inverse PCR with two mutagenic primers oriented in opposite directions and directly flanking K59. After phosphorylating with polynucleotide kinase, these primers were used for PCR amplification of each of 418 pBD-PYR1 mutant templates. Three ligand-contacting residues (P55, F61 and I62) are too close to K59 to utilize this method. To introduce K59R into mutants at these sites, individual K59R mutagenic primers were designed complementary to each of the 57 remaining mutant templates. These primers were then used for inverse PCR mutagenesis, as described earlier. The linear PCR products generated using either method were ligated using T4 DNA ligase, digested with the restriction enzyme Dpn1 (to remove original template DNA) and transformed into competent *Escherichia coli* cells. Transformed colonies were screened by PCR using K59R allele-specific primers to identify plasmids that had successfully incorporated the K59R mutation. K59R mutant plasmids were isolated and sequenced to verify that they contained both the introduced K59R mutation and the original ligand-site mutation. This mutagenesis effort created a set of 475 PYR1(K59R) variants containing all possible single amino acid substitutions at 25 ligand-contacting residues. Sanger sequencing was used to validate all clones constructed.

Assays of the pocket library for agrochemical responsiveness. The set of 475 mutant plasmids were individually transformed into the Y190 yeast two-hybrid reporter strain co-transformed with pACT-HAB1, as previously described[6]. The yeast strains generated were arrayed into 96-well plates yielding what we refer to as the 'pocket library'. The pocket library strains were spotted onto duplicate agar plates containing selective synthetic dextrose (minus Leu and Trp) medium that was supplemented with a single test compound at 100 μM. The pocket library strains were separately tested for responsiveness to the following compounds: benzothiadiazole, mandipropamid, fludioxonil, benoxacor, mesotrione, thiamethoxam, cyprrodinil, azoxystrobin, primicarb, lufenuron, tefluthrin, fomasafen, cloquintocet, fenclorim and cloquintocet-mexyl. All agrochemicals used were purchased from Sigma-Aldrich. After incubating test plates at 30 °C for 2 days, colonies were chloroform lysed and stained to reveal β-galactosidase expression levels, using previously described methods[6]. Mutants displaying responsiveness to the test compound, if present, were identified by virtue of X-gal staining and then subjected to subsequent optimization efforts.

First round optimization by combinatorial mutagenesis. To identify potential additive or synergistic interactions between the mutations that improve receptor function, we constructed combinations of the best variants identified in the first round of screening for mandipropamid, benzothiadiazole, benoxacor and fludioxonil sensitivity. The mutations selected for combinatorial mutagenesis are marked with asterisks in Extended Data Fig. 1. The mutant combinations were constructed using the QuickChange Lightning Multi Site-Directed PCR Mutagenesis kit (Agilent) using pBD-PYR1(K59R) template DNA and mutagenic primers, essentially as previously described[15]. The mutant combinations were sequence validated, introduced into the pACT-HAB1 Y190 reporter strain and then tested for responsiveness to a range of compound concentrations (100, 50, 25, 10, 1, 0.2, 0.1 or 0 μM each test compound). These efforts yielded double-mutant receptor variants with improved sensitivity for the four compound–receptor pairs examined; however, only mandipropamid and benzothiadiazole receptors displayed responses at concentrations as low as 1 μM. Efforts to improve the benzothiadiazole receptor further were abandoned because we could not produce active recombinant protein for biochemical

characterization. The most sensitive response to mandipropamid was obtained with the triple mutant PYR1(K59R/F108A/S122G) (Extended Data Fig. 2).

Site-saturation mutagenesis. We screened for additional pocket-located mutations that would improve the mandipropamid responsiveness of the PYR1(K59R/F108A/S122G) receptor. Extended Data Figure 2A outlines the engineering scheme for developing the receptor and Extended Data Fig. 2B the specific mutants tested for activity. We first conducted site-saturation mutagenesis of 23 pocket-lining residues using NNK primers, which enable all amino acids at a targeted site to be encoded while only encoding one of the three possible stop codons. Each NNK primer was used to generate a pool of PYR1(K59R/F108A/S122G) receptors which were then combined and transformed into a pACT-HAB1 MAV99 reporter strain, which allows URA3-based negative selection against constitutively active receptors and positive selection of activated receptors[17]. The pooled yeast cells were first plated onto synthetic dextrose (without Trp and Leu) media containing 0.1% FOA to select against constitutively active receptors. After 2 days, the surviving yeast cells were plated onto SD (without Ura, Trp and Leu) plates containing 1 μM mandipropamid, a concentration too low to allow growth of a control MAV99 pACT-HAB1 reporter strain expressing pBD-PYR1(K59R/F108A/S122G). Colonies displaying growth on the selective medium were identified and subsequently re-tested on selective media with and without mandipropamid. These efforts yielded ten mutations in five residues (V81C, V81I, V81T, V83L, L87A, F159L, F159M, F159V, A160V, V164I) that enhance the mandipropamid sensitivity of PYR1(K59R/F108A/S122G) (see Extended Data Fig. 2B).

DNA shuffling. We also used recombination-based mutagenesis to identify mutant combinations that enhance PYR1(K59R/F108A/S122G) sensitivity using nucleotide excision and exchange technology (NExT)[25]. An equal amount of PYR1(K59R/F108A/S122G) template was recombined with an equal amount of template DNA that was made by pooling plasmid DNAs from the PYR1(K59R/F108A/S122G) NNK plasmid libraries described earlier. A ~200,000 member library of mutagenized clones was generated and was transformed into the MAV99 pACT-HAB1 reporter strain. Selections were conducted on plates containing 1 μM mandipropamid and characterized for ligand-dependent interactions, as described earlier. These efforts identified Y58H as an additional mutation that improves the mandipropamid sensitivity of the PYR1(K59R/F108A/S122G) receptor. Y58 was not targeted in the PYR1(K59R/F108A/S122G) NNK library and was therefore a spontaneous mutant that arose during the mutagenesis process. The side chain of Y58 projects into PYR1's ligand-binding pocket, but is not within the 5 Å cut-off that we initially employed for targeting pocket residues for site-saturation mutagenesis.

Final optimization using combinatorial mutagenesis. Mutagenic primers for all of the strongest enhancing mutations identified in the NNK-mutagenesis (V81I, V83L, F159L, A160V and V164I) were designed and used simultaneously with the QuickChange Lightning Multi Site-Directed Mutagenesis kit (Agilent) using PYR1(Y58H/K59R/F108A/S122G) template DNA. Individual clones were sequenced to identify combination mutants, which were transformed into the Y190 pACT-HAB1 yeast strain and assayed for mandipropamid sensitivity on varying concentrations of mandipropamid. This led to the identification of PYR1(Y58H/K59R/V81I/S122G/F108A/F159L), PYR1MANDI, which responds to mandipropamid concentrations as low as 10 nM in the yeast assay (Fig. 1b and Extended Data Fig. 2B).

Receptor- and ligand-mediated PP2C inhibition. PYR1MANDI and 9 of the 11 wild-type receptors characterized (previously cloned[20]) were expressed as 6×His-tagged fusion proteins in pET28; PYL9 and PYL11 were expressed as maltose-binding fusion proteins using the pMAL-c expression vector[20]. Recombinant 6×His-tagged receptors[20] and GST–PP2Cs[6] were expressed and purified as previously described[6,20]. PP2C assays were conducted in 96-well polystyrene flat-bottom microtitre plates (Greiner). Assays for all receptors except PYL9 were conducted using the following assay conditions: 100 nM 6×His–receptor, 50 nM GST–PP2C, 100 mM Tris-HCl (pH 7.9), 100 mM NaCl, 1 mM MnCl$_2$, 1% β-mercaptoethanol and 0.3% BSA. Reactions were mixed with probe molecules (or mock carrier solvent-only controls), equilibrated for 30 min, after which 4-methylumbeliferyl phosphate was added (1 mM final concentration). The plates were read using a Victor 2 plate reader (PerkinElmer) (355 excitation, 460 nm emission). PYL9 was assayed at 300 nM, under otherwise identical conditions. GST–HAB1, GST–ABI1 and GST–ABI2 were all tested with identical reactions, which were run in triplicate. The ratio of receptor to PP2C used in our assays was selected based on titration experiments, which showed that maximal inhibition of HAB1 PP2C activity (at saturating ABA concentrations, 10 μM) required a twofold excess of each receptor to PP2C; the total PP2C concentration used (50 nM; established empirically) probably overestimates active PP2C concentrations since we observed some IC$_{50}$ values below 50 nM. PP2C activity values reported are expressed as per cent control values, which were calculated by including the carrier solvent (1% dimethylsulphoxide (DMSO)) and the specific receptor assayed, but no probe molecule.

Receptor–PP2C pull-down assays. PYR1MANDI was cloned as a 6×His–GFP fusion protein in pEGAD[26]. HAB1 was cloned as a GFP fusion protein in the vector pEGAD

(without a 6×His tag). Both constructs were transformed into *Agrobacterium tumefaciens* (GV3101) and these strains and a strain expressing the silencing suppressor p19 were mixed together in ratios corresponding to 0.1 receptor, 1.0 HAB1 and 0.5 p19 final $OD_{600\,nm}$ units, respectively; this yielded roughly equimolar amounts of PP2C and receptor. The mixture was infiltrated into two separate *N. benthamiana* leaves and 2 days later the leaves were treated with either 50 μM mandipropamid or mock solutions made in water containing 0.02% Silwet L-77 (obtained from Lehle Seeds). After 20 h, the leaves were homogenized in liquid nitrogen, resuspended in an extraction buffer composed of 1× TBS, 0.1% NP-40, 1 mM dithiothreitol (DTT), 10% glycerol and 1× plant protease inhibitor cocktail (USB) and clarified by centrifugation. Twenty-five milligrams of PrepEase Ni-TED beads (USB) was added to 2 ml of each extract to isolate 6×His-tagged PYR1[MANDI] receptors and associated proteins. The resin was washed three times; bound proteins were eluted in SDS–PAGE loading buffer, separated by SDS–PAGE and then electroblotted onto nitrocellulose membranes. Both PYR1[MANDI] and HAB1 were expressed as GFP fusion proteins and detected using an anti-GFP monoclonal primary antibody (Clontech) and an anti-mouse IgG-HP sheep secondary antibody (GE Healthcare) using ECL (Perkin Elmer) development.

Protein crystallization. PYR1(K59R/V81I/F108A/F159L) and ΔNHAB1 (residues 1–178 deleted) were expressed in *E. coli* and purified as described previously[6,9]. Purified PYR1(K59R/V81I/F108A/F159L) was mixed with an eqimolar amount of ΔNHAB1 and a fivefold molar excess of mandipropamid and incubated at room temperature for 10 min. The mixture was then exchanged into 20 mM Tris (pH 7.6), 50 mM NaCl solution and concentrated to 15 mg ml^{-1}. Crystallization was conducted at 19 °C by sitting-drop vapour diffusion by mixing equal volumes of the protein with well solution containing 100 mM Bis-Tris propane (pH 7.0), 22.5% PEG 2000 monomethyl ether, and 150 mM sodium malonate. The resulting crystals were flash frozen after passing through a cryoprotection solution of the well solution plus an additional 20% glycerol. All diffraction data were collected at 100 K using an R-AXIS IV^{++} detector equipped with a MicroMax007 generator and an Osmic mirror set. Diffraction data were processed with HKL2000.

X-ray crystal structure determination. Molecular replacement was used to evaluate the initial phases using the PYR1–ABA–HAB1 complex (Protein Data Bank accession 3QN1) as the search model. Phenix.AutoMR solved the initial phases and automatically built in the majority of the residues for the ternary complex. Models were completed through iterative rounds of manual model building in Coot and refinement with Phenix.refine using translational libration screw-motion (TLS) and individual atomic displacement parameters. Mandipropamid was modelled using the ProgDrg server and placed into the complex after several rounds of manual refinement to limit model bias. The geometry of the final structure was validated using Molprobity and Procheck. Ramachandran statistics for the ternary complex were 98.3 and 1.7% for the favoured and additionally allowed regions of the Ramachandran plot, respectively. Data collection and refinement statistics for the final model are listed in Extended Data Table 2.

Production of transgenic *Arabidopsis*. The PYR1[MANDI] coding sequence was PCR amplified from the pBD-PYR1[MANDI] template and cloned into the plant transformation vector pEGAD under control of the 35S promoter. This construct was introduced into *A. tumefacians* GV3101 and then used to transform *Arabidopsis* using the floral dip method[27] and the resultant seed was germinated in soil and treated with glufosinate to identify transformed plants. Seed from approximately 15–16 transgenic plants were harvested individually and used to identify single insert lines. Three independent homozygous single insert 35S::PYR1[MANDI] insertion lines (referred to as lines 1, 2 3 in Extended Data Figs 5 and 8) were used in this work. The RNA-seq experiments used line 1, which had the highest-level expression of PYR1[MANDI] protein, and the drought experiments used both lines 1 and 2. Protein blots comparing protein expression levels in the lines are presented in Extended Data Fig. 4. Western blots characterizing PYR1 protein expression levels used a previously described and validated polyclonal PYR1 antibody[28].

Root growth and seed germination assays. The wild-type, PYR1[OX] and PYR1[MANDI] genotypes were surface sterilized in bleach and plated on to 0.7% agar Petri plates containing one-half MS salts and one-half 0.5% sucrose. After 4 days of stratification at 4 °C, the plates were transferred to a growth chamber in darkness and allowed to germinate for 24 h and then transferred to Petri plates (0.7% agar containing one-half MS salts and one-half 0.5% sucrose) supplemented with differing concentrations of mandipropamid. These plates were then grown vertically under darkness. The amount of new root growth after transfer was measured 72 h after transfer. Seeds for germination assays were prepared similarly except that the plates contained differing concentrations of mandipropamid.

RNA-seq experiments. We examined the effects of ABA and mandipropamid on gene expression in wild type and the PYR1[MANDI] (line 1) transgenic lines. Seed of the wild type or transgenic lines were surface sterilized, stratified for 4 days at 4 °C and then grown for 10 days at room temperature under continuous illumination in a liquid culture consisting of one-half MS salts and one-half 0.5% sucrose and

grown with continuous shaking to provide aeration. After 10 days, the culture solutions were adjusted to contain 50 μM ABA, 2 μM mandipropamid, or a mock treatment. Each compound/mock exposure was conducted in biological triplicate. The (+)-stereoisomer of ABA (Biosynth, AG) was used in this study and mandipropamid (mixed stereoisomer) was purchased from Sigma-Aldrich. After 8 h exposure to the test compounds, RNA was isolated using RNAEasy Plant RNA isolation kit (Qiagen) and treated with DNase. The total RNA was prepared for RNA-seq using the NEBNext platform, which consists of a Poly(A) mRNA Magnetic Isolation Module, NEBNext Multiplex Oligos for Illumina, and NEBNext Ultra RNA Library Prep Kit for Illumina, New England BioLabs. poly(A) mRNA was isolated using NEBNext oligo d(T)$_{25}$ magnetic beads and 5 μg total RNA input, as described by the manufacturer. mRNA was eluted using the kit's first-strand synthesis reaction buffer and hybridized to a random primer mix by incubating the sample at 94 °C for 15 min followed by cooling. First-strand cDNA was synthesized using ProtoScript II reverse transcriptase and subsequently second-strand synthesis reactions were conducted using the kit's components. The double-stranded cDNA produced was purified using Agencourt AMPure XP beads and NEBNext adaptors were ligated to the purified cDNAs. The adaptor-ligated DNA was then size-selected using Agencourt AMPure XP beads. Quantities of the size-selected cDNA were increased by PCR enriched (using the manufacturer's protocol) and purified using AMPure XP beads. Library quality was assessed using a Bioanalyzer. The libraries were sequenced using a HiSeq instrument. Each lane was used to analyse six barcoded libraries. Data from 51 cycles of single-end reads were collected. This ultimately yielded approximately 20 million mapped reads per barcoded sample.

The raw HiSeq data was mapped to the *Arabidopsis* genome using TOPHat, which matches Illumina reads to the *Arabidopsis* reference genome (TAIR 10) sequence and calculates the abundance of each gene using the FPKM (fragments per kilobase of exon per million fragments mapped) metric. The Cufflinks[29] package was used to compare samples and the cuffdiff results (showing mean transcript FPKM values and significant differences for all pairwise comparisons) is available (see Source Data associated with Fig. 3). Twenty-two thousand three-hundred and twenty-six genes displayed non-zero mean FPKM values in all samples and, of these, 20,534 genes had mean FPKM values greater than 0.1 across all samples and these were used to make the comparisons of global gene expression patterns shown in Fig. 3.

Drought stress assays. The wild-type Columbia, 35S::PYR1-GFP (PYR1[OX]) and PYR1[MANDI] genotypes were used for drought stress tolerance assays. Each experiment was conducted independently of the other at different times over the course of 8 months. Lines 1 and 2 are two independent homozygous single-insertion transfer-DNA lines (their construction is described earlier); the PYR1[OX] line, which complements the *pyr1-1* mutant phenotype, has previously been described[6]. Each experiment was initiated with five pots of each genotype (four plants per pot) and entire pots were discarded if any of the seedlings died after transplantation; however, each experimental replicate contained a minimum of three pots (raw data shown in Extended Data Fig. 8). After 2 weeks (experiment 1) or 3 weeks (experiments 2 and 3) growth under standard water regimes, watering was ceased and the plants were treated with a mock or 1 μM mandipropamid solution made in water containing 0.02% Silwet L-77; the plants were treated a second time 4 days (experiment 1) or 3 days (experiments 2 and 3) after the initial treatment. Leaf turgor of the wild-type controls was monitored in each of the three experimental sets to identify endpoints based on wilting (between 10 to 12 days, which varies between the three experiments because of differences in ambient humidity and other growth parameters), at which point all plants in the experimental series were re-watered and then survival rates were assessed by restoration of leaf turgor 24 h after re-watering.

Generation and characterization of transgenic tomato expressing PYR1[MANDI]. The 35S::PYR1[MANDI] construct used for construction of *Arabidopsis* transgenics was modified to contain a kanamycin selection marker. Tomato was transformed by *Agrobacterium*-mediated transformation as described previously[30], with minor modifications. Surface-sterilized tomato seeds (strain UC82B) were germinated on sterilized wet filter paper. Cotyledons from 7-day-old seedlings were excised and dipped into a suspension of *Agrobacterium* in MS medium containing 100 μM acetosyringone and 10 μM 2-mercatoethanol for 10 min. Explants were then placed on co-cultivation medium containing MS salts, 3% sucrose, 0.3% Gelite and 1.5 mg l^{-1} zeatin. After 3 days of co-culture in darkness, the explants were transferred onto callus induction medium containing MS salts, 3% sucrose, 0.3% Gelite, 1.5 mg l^{-1} zeatin, 100 mg l^{-1} kanamycin and 125 mg l^{-1} carbenicillin. Explants were transferred to freshly prepared medium every 2 weeks. Calli-displaying shoot buds were transferred to the medium containing 1 mg l^{-1} zeatin to simulate shoot elongation. Transgenic shoots 1 cm in length were cut and transferred onto rooting medium containing MS salts, 1.5% sucrose, 0.3% Gelite, 1 mg l^{-1} isobutyric acid, 50 mg l^{-1} kanamycin and 125 mg l^{-1} carbenicillin. After 2–3 weeks young plants displaying well-developed roots were transferred to the soil. The experiments shown in Extended Data Fig. 7D were conducted using a primary transgenic line and those shown in

Fig. 4 and Extended Data Fig. 7A, C used T2 progeny from an independent single-insert transgenic line. The T2 seed was germinated on filter paper and transferred to soil. Proteins from leaf tissue were probed with an anti-PYR1 antibody to infer protein expression levels. Plants lacking expression were confirmed by PCR to be non-transgene-containing segregants (that is, null segregants). We propagated the transgenic plants, null segregants and wild-type controls by making cuttings from plants of the same age. To make these clones, ~5-cm-long shoots were excised and planted in soil after treatment with a commercial rooting powder (Bonide). Plants were grown in a growth chamber on a 16 h light cycle at 25 °C. About 3 weeks after clone establishment, thermal images were collected and the plants were then treated with a solution containing 25 μM mandipropamid, 0.1% DMSO and 0.05% Silwet-77. Thermal images were then collected 24 h after treatment. The experiment shown in Extended Data Fig. 7D used clones derived from an independent primary transgenic line and treated with a lower mandipropamid concentration (10 μM). For this experiment, ~3 weeks after clones were established, the transgenic and wild-type controls were treated with a mock solution (0.1% DMSO and 0.05% Silwet-77) and then thermal images were collected after 24 h. Three days later the plants were treated with a solution containing 10 μM mandipropamid, 0.1% DMSO and 0.05% Silwet-77. Thermal images were then taken again 24 h after treatment.

24. DeSantis, G. *et al.* Creation of a productive, highly enantioselective nitrilase through gene site saturation mutagenesis (GSSM). *J. Am. Chem. Soc.* **125,** 11476–11477 (2003).
25. Müller, K. M. *et al.* Nucleotide exchange and excision technology (NExT) DNA shuffling: a robust method for DNA fragmentation and directed evolution. *Nucleic Acids Res.* **33,** e117 (2005).
26. Cutler, S. R., Ehrhardt, D. W., Griffitts, J. S. & Somerville, C. R. Random GFP::cDNA fusions enable visualization of subcellular structures in cells of *Arabidopsis* at a high frequency. *Proc. Natl Acad. Sci. USA* **97,** 3718–3723 (2000).
27. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16,** 735–743 (1998).
28. Nishimura, N. *et al.* PYR/PYL/RCAR family members are major *in-vivo* ABI1 protein phosphatase 2C-interacting proteins in *Arabidopsis*. *Plant J.* **61,** 290–299 (2010).
29. Trapnell, C. *et al.* Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnol.* **28,** 511–515 (2010).
30. Sun, H.-J., Uchii, S., Watanabe, S. & Ezura, H. A highly efficient transformation protocol for Micro-Tom, a model cultivar for tomato functional genomics. *Plant Cell Physiol.* **47,** 426–431 (2006).
31. Laskowski, R. A. & Swindells, M. B. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.* **51,** 2778–2786 (2011).

LETTER

Exit from dormancy provokes DNA–damage–induced attrition in haematopoietic stem cells

Dagmar Walter[1]*, Amelie Lier[1]*, Anja Geiselhart[2], Frederic B. Thalheimer[3], Sina Huntscha[1], Mirko C. Sobotta[4], Bettina Moehrle[5], David Brocks[2], Irem Bayindir[2], Paul Kaschutnig[2], Katja Muedder[6], Corinna Klein[1], Anna Jauch[7], Timm Schroeder[8], Hartmut Geiger[5,9], Tobias P. Dick[4], Tim Holland-Letz[10], Peter Schmezer[11], Steven W. Lane[12], Michael A. Rieger[3], Marieke A. G. Essers[1,13], David A. Williams[14,15,16,17], Andreas Trumpp[1,6] & Michael D. Milsom[1,2]

Haematopoietic stem cells (HSCs) are responsible for the lifelong production of blood cells. The accumulation of DNA damage in HSCs is a hallmark of ageing and is probably a major contributing factor in age-related tissue degeneration and malignant transformation[1]. A number of accelerated ageing syndromes are associated with defective DNA repair and genomic instability, including the most common inherited bone marrow failure syndrome, Fanconi anaemia[2,3]. However, the physiological source of DNA damage in HSCs from both normal and diseased individuals remains unclear. Here we show in mice that DNA damage is a direct consequence of inducing HSCs to exit their homeostatic quiescent state in response to conditions that model physiological stress, such as infection or chronic blood loss. Repeated activation of HSCs out of their dormant state provoked the attrition of normal HSCs and, in the case of mice with a non-functional Fanconi anaemia DNA repair pathway, led to a complete collapse of the haematopoietic system, which phenocopied the highly penetrant bone marrow failure seen in Fanconi anaemia patients. Our findings establish a novel link between physiological stress and DNA damage in normal HSCs and provide a mechanistic explanation for the universal accumulation of DNA damage in HSCs during ageing and the accelerated failure of the haematopoietic system in Fanconi anaemia patients.

To interrogate whether DNA damage could be induced *in vivo* by physiological activation of HSCs out of quiescence, we injected C57BL/6J mice with polyinosinic:polycytidylic acid (pI:pC) to mimic a viral infection and induce a type I interferon response. As previously described[4], administration of pI:pC led to a rapid transition of long-term (LT)-HSCs from quiescence into an active cell cycle *in vivo* (Fig. 1a and Extended Data Fig. 1a–f). Using the alkaline comet assay, we observed increased DNA double-strand breaks (DSBs) and single-strand breaks in LT-HSCs isolated from mice treated with pI:pC, consistent with the hypothesis that HSC quiescence facilitates genomic stability[5] (Fig. 1b, c). This finding was recapitulated when we interrogated the frequency of LT-HSCs that harboured nuclear foci of γH2AX, a surrogate marker of DSBs, and both 53BP1 and RAD51 foci, which comprise downstream repair intermediates of DNA DSBs (Fig. 1d–f). A comparison of the kinetics of γH2AX foci induction suggested that *de novo* DNA damage follows exit out of quiescence (Extended Data Fig. 1g). To test the relationship between cell cycle induction and the DNA damage response (DDR) *in vivo*, we made use of two genetic mouse models that have a perturbed HSC response to pI:pC treatment. HSCs in mice with deletions of either the type I interferon receptor ($Ifnar^{-/-}$) or stem cell antigen-1 ($Sca-1^{-/-}$; $Sca-1$ also known as $Ly6a$), fail to enter the cell cycle in response to pI:pC treatment. However, $Sca-1^{-/-}$ HSCs still maintain the interferon (IFN)-α transcriptional response whereas $Ifnar^{-/-}$ HSCs do not (Extended Data Fig. 1h–j)[4]. In both models, we could observe a complete rescue of the pI:pC-induced DDR in LT-HSCs (Fig. 1g). Mechanistically, this demonstrates that DNA damage is dependent on the effect of IFN-α on LT-HSCs and is probably a consequence of cell cycle entry as opposed to induction of the IFN-α transcriptional response program.

HSCs can be induced out of quiescence *in vivo* using a range of physiological stimuli that are not DNA-damaging agents, such as treatment with IFN-α, granulocyte colony-stimulating factor (G-CSF), thrombopoietin (TPO) or by serial bleeding[4,6–8]. We found that all these physiological stimuli were able to induce a robust cell cycle entry response from LT-HSCs *in vivo* (Extended Data Fig. 2a, b) and all were able to provoke DNA damage in LT-HSCs in a similar range to that observed for pI:pC (Fig. 1h). Thus, the acquisition of DNA damage in LT-HSCs is a uniform consequence of treating mice with a range of stimuli that provoke HSC cycle entry and model physiological stress situations including infection and chronic blood loss.

Mitochondrial reactive oxygen species (ROS) are thought to act as key DNA-damaging agents in the context of replicative stress. In line with the observation that the metabolic activity of mitochondria is higher in proliferating haematopoietic progenitors compared to LT-HSCs at homeostasis[9,10], we found that mitochondrial membrane potential was increased within the LT-HSC compartment itself upon *in vivo* exit from quiescence (Extended Data Fig. 2c, d). Using retrovirally-encoded genetic redox probes[11,12], we could determine that this altered mitochondrial metabolism correlated with a significant increase in the levels of both cytoplasmic H_2O_2 and of the mitochondrial glutathione redox potential in HSCs *in vivo* (Fig. 2a, b and Extended Data Fig. 2e–h). Critically, we could detect elevated levels of the ROS-specific 8-oxo-2′-deoxyguanosine (8-Oxo-dG) lesion on the DNA of activated LT-HSCs (Fig. 2c and Extended Data Fig. 2c). ROS-induced lesions can ultimately precipitate DNA DSBs if unresolved before DNA replication. Thus, combined overexpression of the ROS-detoxifying enzymes manganese superoxide dismutase (SOD2) and catalase rescued the stress-induced increase in DNA DSBs in LT-HSCs, while still maintaining their ability to exit quiescence *in vivo* (Fig. 2d and Extended Data Fig. 2e, f, i–l). This positions ROS as upstream effectors of DNA damage in stress-induced
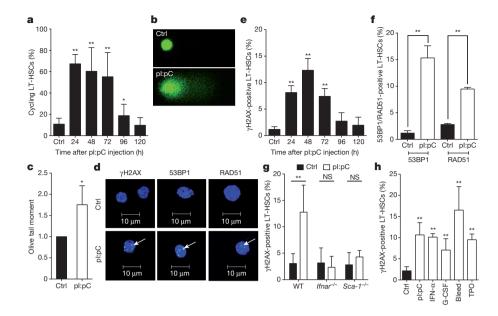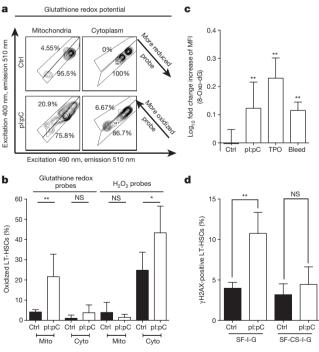
Figure 1 | pI:pC induces proliferation-induced DNA damage in LT-HSCs. a, Time course of LT-HSCs in active cycle (G1/S/G2/M) after pI:pC treatment. Data are mean ± standard deviation (s.d.), $n = 8–11$. **b, c**, Comet analysis of LT-HSCs after pI:pC treatment (**b**) and mean olive tail moment in LT-HSCs (**c**). Data are mean ± s.d., $n = 3$. **d**, Representative LT-HSCs stained with 4′,6-diamidino-2-phenylindole (DAPI; blue) and anti-γH2AX, anti-53BP1 or anti-RAD51 antibodies (green). Arrow indicates foci. **e**, Percentage of γH2AX-positive LT-HSCs after pI:pC treatment. Data are mean ± s.d., $n = 3–4$. **f**, Percentage of 53BP1/RAD51-positive LT-HSCs after pI:pC treatment. Data are mean ± s.d., $n = 3$. **g**, Percentage of γH2AX-positive wild-type (WT), $Ifnar^{-/-}$ or $Sca-1^{-/-}$ LT-HSCs after pI:pC treatment. Data are mean ± s.d., $n = 3–5$. **h**, Percentage of γH2AX-positive LT-HSCs after the indicated treatment. Data are mean ± s.d., $n = 3–5$. Ctrl, control (non-treated). *$P < 0.05$, **$P < 0.01$. Not significant (NS) = $P > 0.05$. Details of exact experimental numbers and statistical analysis can be found in Methods.

LT-HSCs, while more differentiated progenitors are unresponsive to stress and already demonstrate higher homeostatic levels of DNA damage due to their higher proliferative index (Extended Data Fig. 3a–g).

Inactivation of the Fanconi anaemia DNA repair pathway compromises the genomic integrity of HSCs, resulting in bone marrow failure and an increased risk of haematological malignancy[13]. Pro-inflammatory

cytokines have been proposed to act as mediators of the bone marrow failure observed in Fanconi anaemia patients (Extended Data Fig. 4a–c), although the mechanistic link with DNA damage remains unclear[2,14,15]. RNA expression analysis and enumeration of nuclear FANCD2 foci revealed that LT-HSCs engaged a Fanconi anaemia pathway-mediated DDR *in vivo* in response to physiological stress (Fig. 3a, b and Extended Data Fig. 4d–g). In mice harbouring a targeted deletion of *Fanca* ($Fanca^{-/-}$), and that therefore were unable to mount the Fanconi anaemia DDR, administration of pI:pC resulted in higher levels of DNA

Figure 2 | Metabolic ROS induce DNA damage in cycling LT-HSCs.
a, FACS plots showing percentage of LT-HSCs with a predominant oxidized or reduced glutathione redox state after pI:pC treatment. **b**, Percentage of LT-HSCs with a more oxidizing glutathione redox potential or increased endogenous H_2O_2 generation after pI:pC treatment. Cyto, cytoplasm; Mito, mitochondria. Data are mean ± s.d., $n = 5–6$. **c**, Fold change in 8-Oxo-dG lesions within LT-HSCs after the indicated treatment compared to control (Ctrl; non-treated). Mean fluorescent intensity (MFI). Data are mean ± s.d., $n = 4–8$. **d**, Percentage of γH2AX-positive transduced LT-HSCs after pI:pC treatment. SF-CS-I-G, catalase/SOD2-overexpressing vector; SF-I-G, control vector. Data are mean ± s.d., $n = 4–5$. *$P < 0.05$, **$P < 0.01$. Not significant (NS) = $P > 0.05$.
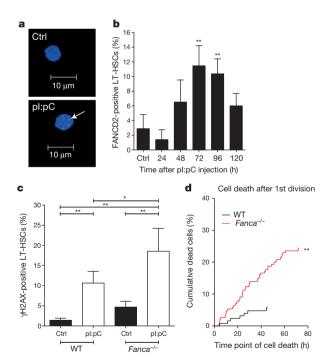

Figure 3 | The Fanconi anaemia pathway is involved in resolving *in vivo* proliferative stress in LT-HSCs. a, Representative LT-HSCs stained with DAPI (blue) and anti-FANCD2 antibody (green) isolated from control (Ctrl; non-treated) or pI:pC-treated mice. Arrow indicates FANCD2 focus. **b**, Percentage of FANCD2-positive LT-HSCs after pI:pC treatment. Data are mean ± s.d., $n = 3–4$. **c**, Percentage of γH2AX-positive LT-HSCs after pI:pC treatment. WT, wild type. Data are mean ± s.d., $n = 3–5$. **d**, Cumulative incidence of LT-HSC death after first *in vitro* division. $n = 126–187$. *$P < 0.05$, **$P < 0.01$.
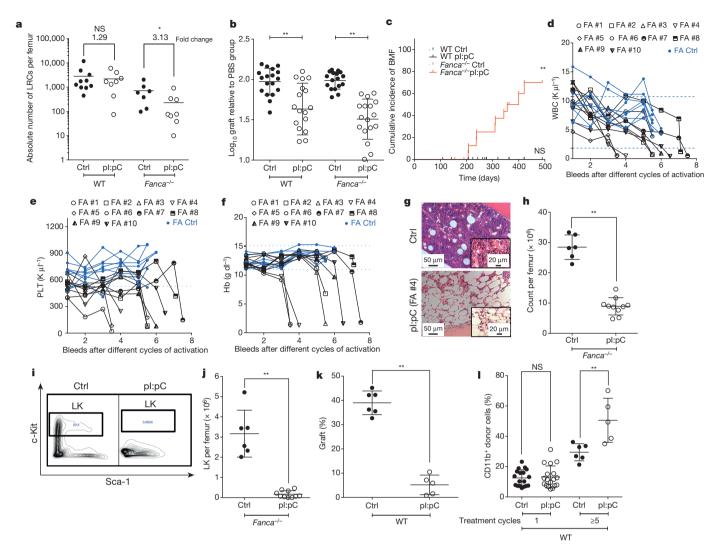
Figure 4 | Chronic stress leads to LT-HSC depletion and bone marrow failure in *Fanca*−/− mice. a, Frequency of bone marrow label-retaining LT-HSCs (LRCs) after an extended pI:pC schedule. WT, wild type. Individual mice plus mean are indicated. $n = 7–10$. **b,** Competitive repopulation assay using bone marrow cells from mice subjected to an extended pI:pC schedule. Individual donor mice, mean ± s.d. are indicated. $n = 17–18$. **c,** Cumulative bone marrow failure (BMF) incidence in wild-type (WT) or *Fanca*−/− mice after repeated pI:pC treatment. $n = 12–18$. **d–f,** Peripheral blood cell analysis from PBS- (blue) or pI:pC- (black) treated *Fanca*−/− mice. Total leukocyte counts (WBC; **d**), platelets (PLT; **e**) and haemoglobin levels (Hb; **f**). Individual bone marrow failure mice are numbered (FA #1 and so on). **g, h,** Representative haematoxylin and eosin (H&E)-stained sections of bone marrow (**g**) and femur cellularity (**h**) in control or pI:pC-treated *Fanca*−/− mice. Data are mean ± s.d., $n = 6–10$. **i,** FACS plots showing frequency of lineage− c-Kit+ (LK) HSCs/progenitors in bone marrow of control or pI:pC-treated *Fanca*−/− mice. **j,** Frequency of LK cells per femur in control or pI:pC-treated *Fanca*−/− mice. Data are mean ± s.d., $n = 6–10$. **k,** Competitive repopulation assay using bone marrow cells from control or pI:pC-treated wild-type mice. Percentage graft at 12 weeks after transplantation is shown. Data are mean ± s.d., $n = 5–6$. **l,** Percentage of CD11b+ donor cells in peripheral blood at 12 weeks after transplantation. Data are mean ± s.d., $n = 5–18$. Control, Ctrl (PBS treated). *$P < 0.05$, **$P < 0.01$. Not significant (NS) = $P > 0.05$.

damage in LT-HSCs compared to their wild-type counterparts (Fig. 3c and Extended Data Fig. 4h, i). These data identify the Fanconi-anaemia-mediated DDR as a route via which proliferative DNA damage is resolved in LT-HSCs *in vivo*. Using an *in vitro* model of HSC activation, we could observe that the *in vivo* stress phenotypes of altered mitochondrial metabolism, increased ROS-induced DNA lesions and elevated DNA DSBs, were all conserved when fluorescence-activated cell sorting (FACS)-isolated LT-HSCs were induced to exit quiescence by *in vitro* culture in the absence of any exogenous pro-inflammatory cytokines (Extended Data Fig. 4j–m). Time-lapse imaging-based single-cell fate tracking[16,17] of LT-HSCs demonstrated high rates of cell death in the absence of a functional Fanconi anaemia DNA repair pathway, even in the very first division after exit from quiescence (Fig. 3d, Extended Data Fig. 4n, o and Supplementary Videos 1, 2). This suggests that inefficient repair of replicative DNA damage may result in LT-HSC depletion.

We next subjected wild-type and *Fanca*−/− mice to an extended 4-week pI:pC treatment regimen to determine whether stress-induced

DNA damage had any prolonged impact upon HSC function (Extended Data Fig. 5a). Four weeks after serial treatment with pI:pC, both wild-type and *Fanca*−/− LT-HSCs had returned to a quiescent state, similar to PBS-treated controls (Extended Data Fig. 5b). However, using an *in vivo* label retention assay (Extended Data Fig. 5c) we could discern that, whereas wild-type LT-HSCs were still able to re-enter a long-term quiescent state after stress-induced cycling, *Fanca*−/− mice demonstrated a threefold reduction in dormant label-retaining LT-HSCs (Fig. 4a). We next performed competitive repopulation experiments to address whether the loss of dormant *Fanca*−/− LT-HSCs correlated with compromised functional activity (Extended Data Fig. 5a). Whereas we could observe a twofold reduction in the repopulating activity of wild-type HSCs after serial pI:pC treatment, there was a fourfold decrease in *Fanca*−/− HSCs, which depleted their absolute frequency to the limit of detection using this assay (Fig. 4b and Extended Data Fig. 5d).

Even though HSCs were functionally compromised in pI:pC-treated *Fanca*−/− and wild-type mice at the time they were euthanized, we

could detect no abnormalities in the levels of mature haematopoietic cells or progenitors (Extended Data Fig. 5e–i). However, after further rounds of HSC activation we found that almost 80% of pI:pC-treated *Fanca*$^{-/-}$ mice succumbed to a severe aplastic anaemia within 3–7 rounds of the 8-week treatment cycle, whereas this was not evident in pI:pC-treated wild-type mice, nor in PBS-treated wild-type or *Fanca*$^{-/-}$ mice (Fig. 4c and Extended Data Fig. 6a). Bone marrow failure was characterized by profound peripheral blood cytopenias accompanied by pronounced bone marrow aplasia (Fig. 4d–h and Extended Data Fig. 6b–j). Megakaryocyte precursors in the bone marrow were depleted, while there was no evidence of myelofibrosis or morphological dysplasia (Extended Data Fig. 6k–m). Flow cytometry analysis of residual bone marrow cells demonstrated the almost complete exhaustion of the entire HSC and progenitor compartment (Fig. 4i, j). Although bone marrow failure was not evident in wild-type mice serially treated with pI:pC, end-point analysis demonstrated a profound depletion of transplantable HSCs, coupled with a strong myeloid lineage bias upon reconstitution of recipient mice (Fig. 4k, l). This suggests that cumulative rounds of replicative stress in wild-type mice can precipitate an HSC phenotype akin to accelerated ageing.

Our data establish a novel connection between stress haematopoiesis and the occurrence of DNA damage and functional decline in adult HSCs. Given the fact that an organism will be exposed to serial rounds of infection and trauma during a lifetime, this provides an attractive explanation for the accumulation of DNA damage in HSCs with age and the associated atrophy within the haematopoietic system (Extended Data Fig. 7)[18]. This link can be formally established from our findings relating to the accelerated ageing disorder, Fanconi anaemia. Although cells from Fanconi anaemia knockout mice have the same DNA repair defect as their human counterparts, they never spontaneously develop the bone marrow failure observed in almost all patients with the disease[19]. For the first time, to our knowledge, we were able to demonstrate that a physiological stress response could drive the complete collapse of the haematopoietic system in almost all analysed *Fanca*$^{-/-}$ mice, faithfully recapitulating the progression of the human disease. These findings not only provide novel insights into the aetiology of this disease, but also highlight the possible role of stress-induced HSC DNA damage in other inherited and acquired bone marrow failure syndromes that have a pro-inflammatory component, as well as in normal ageing[20].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Rossi, D. J., Jamieson, C. H. & Weissman, I. L. Stems cells and the pathways to aging and cancer. *Cell* **132**, 681–696 (2008).
2. Garaycoechea, J. I. & Patel, K. J. Why does the bone marrow fail in Fanconi anemia? *Blood* **123**, 26–34 (2014).
3. Garinis, G. A., van der Horst, G. T., Vijg, J. & Hoeijmakers, J. H. DNA damage and ageing: new-age ideas for an age-old problem. *Nature Cell Biol.* **10**, 1241–1247 (2008).
4. Essers, M. A. *et al.* IFNα activates dormant haematopoietic stem cells *in vivo*. *Nature* **458**, 904–908 (2009).
5. Flach, J. *et al.* Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. *Nature* **512**, 198–202 (2014).
6. Cheshier, S. H., Prohaska, S. S. & Weissman, I. L. The effect of bleeding on hematopoietic stem cell cycling and self-renewal. *Stem Cells Dev.* **16**, 707–718 (2007).
7. Wright, D. E. *et al.* Cyclophosphamide/granulocyte colony-stimulating factor causes selective mobilization of bone marrow hematopoietic stem cells into the blood after M phase of the cell cycle. *Blood* **97**, 2278–2285 (2001).
8. Yoshihara, H. *et al.* Thrombopoietin/MPL signaling regulates hematopoietic stem cell quiescence and interaction with the osteoblastic niche. *Cell Stem Cell* **1**, 685–697 (2007).
9. Simsek, T. *et al.* The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche. *Cell Stem Cell* **7**, 380–390 (2010).
10. Takubo, K. *et al.* Regulation of glycolysis by Pdk functions as a metabolic checkpoint for cell cycle quiescence in hematopoietic stem cells. *Cell Stem Cell* **12**, 49–61 (2013).
11. Gutscher, M. *et al.* Real-time imaging of the intracellular glutathione redox potential. *Nature Methods* **5**, 553–559 (2008).
12. Gutscher, M. *et al.* Proximity-based protein thiol oxidation by H_2O_2-scavenging peroxidases. *J. Biol. Chem.* **284**, 31532–31540 (2009).
13. Geiselhart, A., Lier, A., Walter, D. & Milsom, M. D. Disrupted signaling through the Fanconi anemia pathway leads to dysfunctional hematopoietic stem cell biology: underlying mechanisms and potential therapeutic strategies. *Anemia* **2012**, 265790 (2012).
14. Haneline, L. S. *et al.* Multiple inhibitory cytokines induce deregulated progenitor growth and apoptosis in hematopoietic cells from *Fac*$^{-/-}$ mice. *Blood* **91**, 4092–4098 (1998).
15. Whitney, M. A. *et al.* Germ cell defects and hematopoietic hypersensitivity to gamma-interferon in mice with a targeted disruption of the Fanconi anemia C gene. *Blood* **88**, 49–58 (1996).
16. Thalheimer, F. B. *et al.* Cytokine-regulated GADD45G induces differentiation and lineage selection in hematopoietic stem cells. *Stem Cell Reports* **3**, 34–43 (2014).
17. Rieger, M. A., Hoppe, P. S., Smejkal, B. M., Eitelhuber, A. C. & Schroeder, T. Hematopoietic cytokines can instruct lineage choice. *Science* **325**, 217–218 (2009).
18. Beerman, I., Seita, J., Inlay, M. A., Weissman, I. L. & Rossi, D. J. Quiescent hematopoietic stem cells accumulate DNA damage during aging that is repaired upon entry into cell cycle. *Cell Stem Cell* **15**, 37–50 (2014).
19. Kutler, D. I. *et al.* A 20-year perspective on the International Fanconi Anemia Registry (IFAR). *Blood* **101**, 1249–1256 (2003).
20. Young, N. S. Pathophysiologic mechanisms in acquired aplastic anemia. *Hematology Am. Soc. Hematol. Educ. Program* **2006**, 72–77 (2006).

Author Contributions D.W., A.L., P.S., S.W.L., A.J., T.S., H.G., T.P.D., M.A.R., M.A.G.E., D.A.W., A.T. and M.D.M. designed experiments and analysed experimental data. D.W., A.L., A.G., F.B.T., M.C.S., B.M., S.H., D.B., I.B., P.K., K.M., P.S., C.K. and M.D.M. performed experiments. T.H.-L. performed statistical analysis on experimental data and S.W.L. provided interpretation of mouse histology samples. D.W., A.L., S.W.L., M.A.G.E., P.S., T.P.D., M.A.R., D.A.W., A.T. and M.D.M. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.D.M. (m.milsom@hi-stem.de).

METHODS

Animals. Mice were housed in individually ventilated cages in the DKFZ animal facility and all experimental procedures were approved by the Animal Care and Use Committees of the German Regierungspräsidium Karlsruhe für Tierschutz und Arzneimittelüberwachung. Wild-type mice were obtained from Harlan Laboratories or Charles River Laboratories (C57BL/6J or B6.SJL-Ptprc[a] Pepc[b]/BoyJ). $Fanca^{-/-}$, $Sca\text{-}1^{-/-}$, $Ifnar^{-/-}$, H2B-GFP and ScltTA mice have all been previously described[21–25]. Both wild-type and $Fanca^{-/-}$ mice were crossed with H2B-GFP and ScltTA mice to perform an assay as described previously[26]. Mice were 8–16 weeks of age at the point at which experimental studies were initiated.

To provoke in vivo cycling of HSCs, mice were injected with either $5\,mg\,kg^{-1}$ pI:pC intraperitoneally (i.p.) (InvivoGen), 2×10^5 units recombinant murine IFN-α subcutaneously (s.c.) (Miltenyi), $450\,\mu g\,kg^{-1}$ recombinant human G-CSF s.c. (Amgen), or $200\,\mu g\,kg^{-1}$ recombinant murine TPO i.p. (Peprotech).

Serial bleeding was performed by puncturing the craniofacial capillary bed and allowing the loss of at least $300\,\mu l$ of peripheral blood. This procedure was performed on days 0, 3, 6 and 9 and the mice were euthanized on day 10. After each bleed, mice were injected s.c. with a volume of sterile PBS that was equivalent to the amount of peripheral blood taken.

For the LRC assay, doxycyline treatment was performed by supplementing the drinking water of experimental mice with $2\,mg\,ml^{-1}$ doxycyline citrate (Sigma).

Peripheral blood cell numbers were evaluated using a Hemavet 950 FS veterinary blood cell counting machine (Drew Scientific).

Identification and purification of murine haematopoietic stem and progenitor cells. Murine lineage-depleted bone marrow cells were isolated essentially as described previously[27]. Briefly, bone marrow cells were harvested from the femora, tibiae, ilia and vertebrae of 8–10-week-old mice by gentle crushing in Iscove's modified Dulbecco's medium (Life Technologies). Low-density mononuclear cells (LDMNCs) were purified by density gradient centrifugation using Histopaque 1083 (Sigma-Aldrich). LDMNCs were then stained with a panel of rat anti-mouse biotin-conjugated lineage markers (see Supplementary Table 1). The labelled LDMNCs were subsequently incubated with Biotin Binder Dynabeads (Life Technologies) and the lineage-positive cells were depleted using a Dynamag-15 magnet (Life Technologies). The resulting lineage-depleted cells were then stained with a panel of antibodies (see Supplementary Table 1) to identify or prospectively isolate LT-HSCs (lineage negative (lin$^-$), c-Kit$^+$, Sca-1$^+$, CD48$^-$, CD150$^+$, CD34$^-$), ST-HSCs (lin$^-$, c-Kit$^+$, Sca-1$^+$, CD48$^-$, CD150$^+$, CD34$^+$) and MPPs (lin$^-$, c-Kit$^+$, Sca-1$^+$, CD48$^+$, CD150$^-$) cell fractions by flow sorting or flow cytometry analysis. All sorting experiments were performed using a BD FACSAria I, II or III flow cytometer (BD Biosciences) according to the gating strategy shown in Extended Data Fig. 1b.

Competitive repopulation assays. Lethally irradiated wild-type CD45.2 recipient mice (10 Gy; TBI, Bestrahlungsgerät/Buchler GmbH, caesium source) were co-injected intravenously (i.v.) with a mixture of 2×10^6 wild-type CD45.1/CD45.2 whole bone marrow competitor cells and either 2×10^6 wild-type CD45.1 whole bone marrow cells or 6×10^6 $Fanca^{-/-}$ CD45.1 whole bone marrow cells. Each individual recipient mouse was the recipient of bone marrow cells from one individual test mouse, in addition to the competitor bone marrow cells, which were aliquoted from a pool that was common for the whole experimental repeat. The relative frequency of injected material for each recipient mouse was validated by FACS analysis of the mixed input bone marrow cells stained with anti-CD45.1 and anti-CD45.2 antibodies. The repopulating potential of the test bone marrow populations was evaluated by FACS analysis of peripheral blood at 6 months (24 weeks)[27] post-transplantation after staining with CD45.1 and CD45.2 antibodies.

Culturing of LT-HSCs and live cell imaging. LT-HSCs were purified as described earlier and then cultivated in a serum-free medium (StemSpan SFEM) containing 1% penicillin/streptomycin, 2% L-glutamine, $100\,ng\,ml^{-1}$ TPO and $100\,ng\,ml^{-1}$ SCF. After 36 h in culture, the DNA damage was assessed by staining the LT-HSCs for 53BP1, RAD51, γH2AX, 8-Oxo-dG, FANCD2 and Mitotracker as described later.

Time-lapse imaging. FACS-sorted LT-HSCs were resuspended in SFEM containing $100\,ng\,ml^{-1}$ SCF and $100\,ng\,ml^{-1}$ TPO and seeded in 24-well plates equipped with silicon culture inserts (IBIDI). Plates were gas-tight sealed with adhesive tape after 5% CO_2 saturation and immediately subjected to time-lapse imaging. Microscopy was performed using a CellObserver (Zeiss) at 37 °C. Phase contrast images were acquired every 2–3 min using a ×10 phase contrast objective (Zeiss) and an AxioCamHRm camera (at 1,388 × 1,040 pixel resolution) with a self-written VBA module remote controlling Zeiss AxioVision 4.8 software.

Cell tracking. Cell tracking was performed as described[17], until the fate of all progeny in the third generation was determined.

Colony-forming unit assays. Colony-forming unit (c.f.u.) assays were performed by plating whole bone marrow cells into Methocult GF M3434 (Stem Cell Technologies). The number of total colonies was scored after 7 days of culture at 37 °C, 5% CO_2.

Transduction of lineage$^-$, Sca-1$^+$, c-Kit$^+$ cells. The SF91-IRES-eGFP vector has been previously described[27]. The SF91-Cat/SOD2-IRES-eGFP vector for combined overexpression of mitochondrial SOD2 and cytoplasmic catalase was provided by T. Southgate. The Grx1-roGFP2, Mito-Grx1-roGFP2, roGFP2-Orp1 and Mito-roGFP2-Orp1 cassettes[28,29] were each individually sub-cloned into the SF91 vector backbone. Retroviral supernatant was generated by transient calcium phosphate transfection of the phoenix-GP cell line (Nolan Laboratory) as described previously[27]. The transduction of murine lineage$^-$, Sca-1$^+$, c-Kit$^+$ (LSK) cells was performed essentially as described previously[27]. LSKs were isolated and were then cultured in StemSpan SFEM (Stem Cell Technologies, 09650) culture medium containing 2% penicillin/streptomycin (Sigma, P4458; 100 ml), 1% glutamine (Gibco, 25D30-024), $50\,ng\,ml^{-1}$ TPO (Peprotech, 315-14), $50\,ng\,ml^{-1}$ SCF (Peprotech, 250-03) and $50\,ng\,ml^{-1}$ Ftl3 (Peprotech, 300-19) for 2 days on low attachment plates (Cellstar, suspension culture plate) at 37 °C, 5% CO_2. On day 3, cells were transduced at a multiplicity of infection of 40 on cell culture plates pre-coated with $4\,\mu g\,cm^{-2}$ Retronectin (TaKaRa T100B) and with the addition of fresh cytokine-supplemented media. On day 5, cells were harvested and eGFP$^+$ cells were purified by flow sorting. Lethally irradiated (10 Gy) recipient mice were then injected i.v. with $\geq 1 \times 10^5$ gene-modified cells.

FACS analysis. All FACS samples were analysed with a LSRII or LSR-Fortessa cytometer. Dead cells were excluded by using 7-amino actinomycin D (7AAD, Invitrogen).

Ki-67 and Hoechst staining. Bone marrow cells were stained with antibodies directed against cell surface markers as detailed in Supplementary Table 1. After staining, cells were lysed using ACK lysing buffer (Lonza, 10-548E), washed with PBS and then fixed with BD Cytofix/Cytoperm (BD Bioscience) for 20 min at 4 °C. After fixation, cells were washed twice with PermWash (BD Bioscience), re-suspended in 100 μl PermWash containing anti-Ki-67 and incubated overnight at 4 °C. Before analysis, bone marrow cells were incubated with Hoechst 33342 (1/400 diluted) for 10 min at 4 °C.

MitoTracker analysis. Bone marrow cells were incubated with a 50 nM solution of Mitotracker Red CMXRos substrate (Invitrogen) for 30 min at 37 °C. After washing the cells twice with PBS, cells were stained with HSC cell surface markers as detailed in Supplementary Table 1. Cells were washed twice with PBS and then FACS analysed. MFIs were determined based upon the analysis of a minimum of 100 LT-HSCs.

8-Oxo-dG staining. Unless otherwise stated, 8-Oxo-dG staining was performed on bone marrow cells harvested from mice 24 h post-treatment. Bone marrow cells were stained with antibodies against cell surface markers as detailed in Supplementary Table 1, then fixed twice with Cytofix/Cytoperm and Cytofix/Cytoperm plus (BDBioscience). Cells were then treated with DNase for 1 h at 37 °C, then re-suspended in PermWash containing anti-8-Oxo-dG (1/500 diluted, Trevigen, 4354-MC-050) and incubated overnight at 4 °C. After washing, cells were incubated with a secondary antibody (goat anti-mouse Alexa Fluor 488, Invitrogen A-10680) for 1 h at room temperature and then analysed. MFIs were determined based upon the analysis of a minimum of 100 LT-HSCs.

Megakaryocyte and committed progenitor staining. Bone marrow cells were stained with antibodies against cell surface markers as detailed in Supplementary Table 1 and then FACS analysed.

Peripheral blood staining. Thirty microlitres of peripheral blood was stained with antibodies against cell surface markers as detailed in Supplementary Table 1. After red blood cells lysis with ACK lysing buffer, cells were additionally washed with PBS and then FACS analysed.

Annexin V staining of transduced LSK cells. Transduced LSKs were treated for 24 h with 1 mM hydrogen peroxide. After harvest, cells were incubated for 15 min with annexin V according to the manufacturer's instructions (Invitrogen) and then FACS analysed after adding 7AAD.

Redox analysis of transduced LT-HSCs. Transduced bone marrow cells were harvested from mice 24 h after treatment with either PBS or pI:pC and were placed directly on ice. Cells were then stained with a cocktail of antibodies used to prospectively identify LT-HSCs as described in Supplementary Table 1. An aliquot of cells was then treated with 5 mM diamide or 5 mM dithiothreitol (DTT) to force each genetic probe into its completely oxidized or completely reduced state, respectively. Cells were then analysed by flow cytometry and diamide/DTT-treated samples were used to define FACS gates representing oxidized or reduced probe redox states (see Extended Data Fig. 2d). roGFP2-based redox probes are sequentially excited at 400 nm and 490 nm. The fluorescence intensity ratio measured at the emission wavelength of 510 nm defines the redox state of the probe. For detailed information see refs 28, 29.

Immunofluorescence and confocal microscopy. Unless otherwise stated, mice were euthanized at either 24 h (Bleed), 36 h (G-CSF) or 48 h (pI:pC, IFN-α and TPO) after treatment and bone marrow cells were then harvested. Purified LT-HSCs were transferred onto polylysine-coated slides (Thermo Scientific) and fixed for 15 min in 4% paraformaldehyde in PBS. Cells were then permeabilized in 0.2%

Triton-X-100/PBS and incubated for 1 h in TNBS blocking buffer (1% FCS/0.1% Triton-X-100/PBS).

For γH2AX staining, to enumerate a DNA repair intermediate that is formed downstream of a DNA DSB, cells were incubated with rabbit anti-mouse phospho-histone H2A.X (Ser 139) antibody conjugated with Alexa Fluor 488 (20E3, Cell Signaling) diluted 1:200 in 1% FCS/PBS and incubated in a humidified chamber overnight.

FANCD2 staining, to quantify a DNA repair intermediate that indicates activation of the Fanconi anaemia DNA repair pathway, was performed with purified polyclonal rabbit anti-mouse FANCD2 antibody provided by K. J. Patel, which was diluted 1:600 in 1% FCS/PBS. Subsequent staining with a goat anti-rabbit IgG Alexa Flour 488-conjugated secondary antibody (Invitrogen) was performed in a humidified chamber for 60 min at room temperature at a dilution of 1:600 in 1% FCS/PBS.

For 53BP1 and RAD51 staining, to detect DNA repair intermediates downstream of a DNA DSB, cells were first fixed for 10 min in methanol at −20 °C. Cells were then permeabilized in 0.2% Triton-X-100/PBS and blocked for 1 h in TBS-DT (20 mM Tris-HCl, 127 mM NaCl, 0.1% Tween-20, 5% skim milk). Both stainings were performed in a humidified chamber overnight at 4 °C at an antibody dilution of 1:200 for 53BP1 (Novus Biologicals, NB100-305) and 1:100 for RAD51 (Abcam, ab63801). Subsequent staining with a goat anti-rabbit IgG Alexa Fluor 488-conjugated secondary antibody (Invitrogen) was performed in a humidified chamber for 60 min at room temperature at a dilution of 1:600 in TBS-DT.

Between staining steps, slides were washed three times for 5 min in PBS. Cells were counterstained and mounted with ProLong Gold antifade reagent with DAPI (Life Technologies). For negative controls, isotype-matched rabbit IgGs were used in place of primary antibodies.

Images were obtained using a confocal fluorescent microscope (LSM700 from Zeiss). Slides were manually assessed for γH2AX, FANCD2, 53BP1 or RAD51 foci and cells with five or more foci were scored as positive. For each biological repeat, at least 200 individual LT-HSCs were scored.

Alkaline comet assay. Mice were euthanized 48 h after pI:pC treatment and bone marrow cells were harvested. Comet assays were conducted as previously described[30,31], with some modifications. Sorted LT-HSCs were mixed with 0.7% low-melting-temperature agarose (SeaKem), plated on slides (Trevigen) and subsequently lysed overnight. As a positive control, DNA damage was introduced by irradiation of embedded cell aliquots with 5 Gy using a ^{137}Cs radiation source with a dose rate of 1 Gy min^{-1}. Analysis and evaluation of DNA damage was performed by fluorescence microscopy using a fully automated cell scanning system Metafer-4 (Meta-systems) as described previously[32].

Quantitative real-time PCR. The extraction and purification of total RNA from HSCs was performed using the Arcturus PicoPure RNA Isolation Kit (Applied Biosystems) combined with RNase-Free DNase (Quiagen) treatment according to the manufacturer's protocol. cDNA was synthesized using the SuperScript VILO cDNA Synthesis Kit (Invitrogen) according to the manufacturer's protocol. qRT–PCR was performed and analysed on ViiA7 (Applied Biosystems) using Power SYBR Green PCR Master Mix. A list of primers and amplification conditions for each PCR reaction is provided in Supplementary Table 2. Data for each target gene were normalized against both *Oaz1* and *Sdha*.

Western blotting. Whole cell lysates from transduced LSKs were prepared using RIPA buffer (Cell Signaling), 1 mM PMSF (Sigma), 1 mM EDTA, Halt Protease/Phosphatase Inhibitor Cocktail (Pierce) and 4× Roti-Load1 (Roth). Protein lysates were resolved on 4–12% Bis/Tris NuPage gels with MOPS running buffer (Life Technologies) and blotted on nitrocellulose membranes (Amersham International). Membranes were blocked for 1 h in TBS containing 0.1% (v/v) Tween-20 with 20% (w/v) non-fat dry milk powder. Catalase (1/1,000; New England Biolabs, catalogue no. 8841) and manganese SOD (1/1,000; Enzo, ADI-SOD-111-F) primary antibodies were incubated overnight at 4 °C in blocking solution. Secondary antibodies (1/10,000 dilution in blocking solution, Cell Signaling, anti-rabbit-IgG horseradish peroxidase (HRP)-linked antibody, catalogue no. 7074) were incubated for 1 h at room temperaure. Membranes were washed in TBS-Tween 0.1% and immunocomplexes were detected using the ECL kit (Amersham International). Vinculin expression was used as a loading control.

Statistical analysis. Sample sizes for each data set within the main figures were as follows. Fig. 1a: $n = 8$ mice (96 h); $n = 9$ mice (72 h, 120 h); $n = 10$ mice (48 h); $n = 11$ mice (control, 24 h). Fig. 1c: $n = 3$ biological repeats. Fig. 1e: $n = 3$ biological repeats (control, 24 h, 48 h, 72 h, 96 h); $n = 4$ biological repeats (120 h). Fig. 1f: $n = 3$ biological repeats. Fig. 1g: $n = 3$ biological repeats (*Ifnar*$^{-/-}$ control/pI:pC, *Sca-1*$^{-/-}$ pI:pC); $n = 4$ biological repeats (*Sca-1*$^{-/-}$ control); $n = 5$ biological repeats (wild-type/control/pI:pC). Fig. 1h: $n = 3$ biological repeats (IFN-α, G-CSF, TPO); $n = 4$ biological repeats (Bleed); $n = 5$ biological repeats (control, pI:pC). Fig. 2b: $n = 5$ mice (mitochondrial gluatathione control/pI:pC and cytoplasmic H$_2$O$_2$ control/pI:pC); $n = 6$ mice (cytoplasmic glutathione control/pI:pC and mitochondrial H$_2$O$_2$ control/pI:pC). Fig. 2c: $n = 4$ mice (Bleed); $n = 7$ mice

(pI:pC); $n = 8$ mice (control, TPO). Fig. 2d: $n = 4$ biological repeats (control); $n = 5$ biological repeats (pI:pC). Fig. 3b: $n = 3$ biological repeats (120 h); $n = 4$ biological repeats (control, 24 h, 48 h, 72 h, 96 h). Fig. 3c: $n = 3$ biological repeats (*Fanca*$^{-/-}$); $n = 5$ biological repeats (wild type). Fig. 3d: $n = 126$ cells tracked over three independent biological repeats (wild type); $n = 187$ cells tracked over three independent biological repeats (*Fanca*$^{-/-}$). Fig. 4a: $n = 7$ (*Fanca*$^{-/-}$ control); $n = 8$ (wild-type pI:pC, *Fanca*$^{-/-}$ pI:pC); $n = 10$ (wild-type control). Fig. 4b: $n = 17$ (*Fanca*$^{-/-}$ mice); $n = 18$ (wild-type mice). Fig. 4c: $n = 12$ (wild-type/*Fanca*$^{-/-}$ control group); $n = 18$ (wild-type/*Fanca*$^{-/-}$ pI:pC group). Fig. 4h: $n = 6$ (control); $n = 10$ (pI:pC). Fig. 4j: $n = 6$ (control); $n = 10$ (pI:pC). Fig. 4k: $n = 5$ (pI:pC); $n = 6$ (control). Fig. 4l: $n = 5$ (≥5 rounds of treatment cycles; pI:pC group); $n = 6$ (≥5 rounds of treatment cycles; control group); $n = 18$ (1 round of treatment cycle, control/pI:pC).

Unless indicated otherwise in the figure, analyses were always carried out in comparison to the control group. Two-sided *t*-tests were used for two-group comparisons (Figs 1c, f–h, 2b, d, 4a, h, j, k, l and Extended Data Figs 2b, 3b, c, d, e, g, 4g, h, k, l, 5d, 6k) and one- and two-way ANOVAs for comparisons involving more than two groups or more than two independent variables (Figs 1a, e, g, 3b, c and Extended Data Figs 1f, i, 2k, l, 3a, 4e, f, i, 5b, e). For the comparison of HSC frequencies (Fig. 4b) a linear regression with two variables was used and, for the incidence of bone marrow failure considerations in Fig. 4c a Cox regression was performed. For the cell fate tracking experiments in Fig. 3d and Extended Data Fig. 4o, a log-rank Mantel–Cox test was performed. For variables that did not show an at least approximate normal distribution and equal variances (fold change/relative to control data/absolute number, Figs 2c, 4a, b and Extended Data Figs 1j, 2d, 3f, 4j, m) a log$_{10}$ transformation was applied and resulted in reasonably normal data.

Sample size considerations. For pooled experiments requiring the combination of material from several mice (Figs 1e, f, g, h, 2d, 3b, c and Extended Data Figs 1e, j, 4e, f, g, h, i, k) and cell fate tracking (Fig. 3d and Extended Data Fig. 4n, o), three replicates were performed to allow basic statistical inference while using a still justifiable number of animals. For other *in vivo* experiments (Figs 2b, c, 4a, c–f, h, j, k, l and Extended Data Figs 2d, 4g, i, 6k), we generally aimed for a number of six animals per group after all losses. Given a standard deviation of approximately 30% for our log values (based on previous experiences), this allowed us to find changes of 60% with a power of >80% in a two-sided *t*-test, α = 5%. In some situations, observations on additional variables could be taken on mice already available from previous experiments. Specifically, the cell cycle data (Figs 1a, f and Extended Data Figs 2b, k, 4h) could be taken from the pool preparations (before pooling) and Fig. 4a, b included mice from several previous experiments. Therefore, in these cases, larger sample sizes were achieved.

Data exclusions. In Fig. 4c, mice were excluded when the timing of death prevented end-point analysis to confirm a diagnosis of severe aplastic anaemia. A total of 4 wild-type and 2 *Fanca*$^{-/-}$ mice were excluded according to this criterion.

Randomization. Allocation of mice to groups was not formally randomized. However, the possible confounders 'experimenter' and 'day of experiment' were equally matched between groups.

Blinding. No experiments were blinded.

Acquisition and handling of human patient data. The data contained in Extended Data Fig. 4a–c are not considered human subject research by the Boston Children's Hospital Institutional Review Board (IRB) and therefore there is no IRB jurisdiction for IRB review nor any consent requirements.

21. Bockamp, E. *et al.* Tetracycline-controlled transgenic targeting from the *SCL* locus directs conditional expression to erythrocytes, megakaryocytes, granulocytes, and c-kit-expressing lineage-negative hematopoietic cells. *Blood* **108,** 1533–1541 (2006).
22. Müller, U. *et al.* Functional role of type I and type II interferons in antiviral defense. *Science* **264,** 1918–1921 (1994).
23. Stanford, W. L. *et al.* Altered proliferative response by T lymphocytes of Ly-6A (Sca-1) null mice. *J. Exp. Med.* **186,** 705–717 (1997).
24. Tumbar, T. *et al.* Defining the epithelial stem cell niche in skin. *Science* **303,** 359–363 (2004).
25. Wong, J. C. *et al.* Targeted disruption of exons 1 to 6 of the Fanconi anemia group A gene leads to growth retardation, strain-specific microphthalmia, meiotic defects and primordial germ cell hypoplasia. *Hum. Mol. Genet.* **12,** 2063–2076 (2003).
26. Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135,** 1118–1129 (2008).
27. Milsom, M. D. *et al.* Ectopic HOXB4 overcomes the inhibitory effect of tumor necrosis factor-α on Fanconi anemia hematopoietic stem and progenitor cells. *Blood* **113,** 5111–5120 (2009).
28. Meyer, A. J. & Dick, T. P. Fluorescent protein-based redox probes. *Antioxid. Redox Signal.* **13,** 621–650 (2010).
29. Morgan, B., Sobotta, M. C. & Dick, T. P. Measuring E_{GSH} and H$_2$O$_2$ with roGFP2-based redox probes. *Free Radic. Biol. Med.* **51,** 1943–1951 (2011).

30. Schmezer, P. *et al.* Rapid screening assay for mutagen sensitivity and DNA repair capacity in human peripheral blood lymphocytes. *Mutagenesis* **16,** 25–30 (2001).
31. Greve, B. *et al.* Evaluation of different biomarkers to predict individual radiosensitivity in an inter-laboratory comparison—lessons for future studies. *PLoS ONE* **7,** e47185 (2012).
32. Schunck, C., Johannes, T., Varga, D., Lorch, T. & Plesch, A. New developments in automated cytogenetic imaging: unattended scoring of dicentric chromosomes, micronuclei, single cell gel electrophoresis, and fluorescence signals. *Cytogenet. Genome Res.* **104,** 383–389 (2004).

LETTER

Mitochondrial DNA stress primes the antiviral innate immune response

A. Phillip West[1], William Khoury-Hanold[2], Matthew Staron[2], Michal C. Tal[2]†, Cristiana M. Pineda[1], Sabine M. Lang[1], Megan Bestwick[1]†, Brett A. Duguay[3], Nuno Raimundo[1]†, Donna A. MacDuff[4], Susan M. Kaech[2,5], James R. Smiley[3], Robert E. Means[1], Akiko Iwasaki[2,5] & Gerald S. Shadel[1,6]

Mitochondrial DNA (mtDNA) is normally present at thousands of copies per cell and is packaged into several hundred higher-order structures termed nucleoids[1]. The abundant mtDNA-binding protein TFAM (transcription factor A, mitochondrial) regulates nucleoid architecture, abundance and segregation[2]. Complete mtDNA depletion profoundly impairs oxidative phosphorylation, triggering calcium-dependent stress signalling and adaptive metabolic responses[3]. However, the cellular responses to mtDNA instability, a physiologically relevant stress observed in many human diseases and ageing, remain poorly defined[4]. Here we show that moderate mtDNA stress elicited by TFAM deficiency engages cytosolic antiviral signalling to enhance the expression of a subset of interferon-stimulated genes. Mechanistically, we find that aberrant mtDNA packaging promotes escape of mtDNA into the cytosol, where it engages the DNA sensor cGAS (also known as MB21D1) and promotes STING (also known as TMEM173)–IRF3-dependent signalling to elevate interferon-stimulated gene expression, potentiate type I interferon responses and confer broad viral resistance. Furthermore, we demonstrate that herpesviruses induce mtDNA stress, which enhances antiviral signalling and type I interferon responses during infection. Our results further demonstrate that mitochondria are central participants in innate immunity, identify mtDNA stress as a cell-intrinsic trigger of antiviral signalling and suggest that cellular monitoring of mtDNA homeostasis cooperates with canonical virus sensing mechanisms to fully engage antiviral innate immunity.

To explore the cellular responses to mtDNA stress in the absence of oxidative phosphorylation deficiency, we employed a TFAM heterozygous knockout ($Tfam^{+/-}$) mouse model. Cells and tissues from these animals exhibit modest or no significant differences in mtDNA-encoded transcripts and oxygen consumption rates, despite an approximately 50% depletion of mtDNA (Extended Data Fig. 1a–c)[5,6]. In addition to mtDNA depletion, $Tfam^{+/-}$ mouse embryonic fibroblasts (MEFs) have reduced oxidative mtDNA damage repair capacity and markedly altered mtDNA packaging, organization and distribution (Fig. 1a)[6]. Nucleoids in $Tfam^{+/-}$ MEFs were less numerous and exhibited a larger size distribution (Fig. 1a and Extended Data Fig. 1d). Thus, $Tfam^{+/-}$ cells provide a robust model to characterize cellular responses triggered by moderate mtDNA stress.

Gene expression profiling of $Tfam^{+/-}$ MEFs revealed an unexpected enrichment of interferon-stimulated genes (ISGs) and antiviral signalling factors (Fig. 1b). Of the 45 most overexpressed genes, 39 were ISGs, including many with direct antiviral activity ($Ifi44$, $Ifit1$, $Ifit3$, $Oasl2$, $Rtp4$)[7,8]. We also observed increased expression of cytoplasmic RNA and DNA sensors, such as $Ddx58$ and $Ifih1$ and p200 family proteins $Ifi203$, $Ifi204$ and $Ifi205$, as well as transcription factors $Irf7$, $Stat1$ and $Stat2$, ISGs that function to positively reinforce the antiviral response. Direct measurement of basal ISG mRNA and protein expression

in $Tfam^{+/-}$ MEFs validated the microarray results (Fig. 1c, d). Finally, $Tfam^{+/-}$ MEFs expressed three- to fourfold more $Ifnb$ and $Ifna4$ upon transfection with the IFIH1 agonist poly(I:C) (Fig. 1e), consistent with enhanced type I interferon responses.
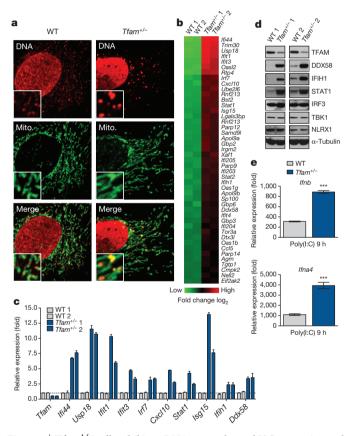

Figure 1 | $Tfam^{+/-}$ cells exhibit mtDNA stress, elevated ISG expression and augmented type I interferon responses. a, Confocal microscopy images of MEFs stained with anti-DNA (DNA) and anti-HSP60 (Mito.) antibodies. **b**, Heat maps of microarray analyses. Genes in $Tfam^{+/-}$ MEFs exhibiting statistically significant ($P < 0.05$), twofold or greater increases over wild type (WT) are shown. **c, d**, Quantitative real-time-PCR (qRT–PCR) (**c**) and western blots (**d**) of basal ISG expression in two littermate wild-type and $Tfam^{+/-}$ MEF lines. **e**, qRT–PCR analysis of type I interferon expression in MEFs 9 h after cytosolic delivery of poly(I:C). Error bars indicate ±s.e.m. of triplicate technical replicates and are representative of three independent experiments. ***$P < 0.001$.

[1]Department of Pathology, Yale School of Medicine, New Haven, Connecticut 06520, USA. [2]Department of Immunobiology, Yale School of Medicine, New Haven, Connecticut 06520, USA. [3]Li Ka Shing Institute of Virology, Department of Medical Microbiology, University of Alberta, Edmonton, Alberta T6G 2S2, Canada. [4]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [5]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815-6789, USA. [6]Department of Genetics, Yale School of Medicine, New Haven, Connecticut 06520, USA. †Present addresses: Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, California 94305, USA (M.C.T.); Department of Chemistry, Linfield College, McMinnville, Oregon 97128, USA (M.B.); Institute for Cellular Biochemistry, Universitätsmedizin Göttingen, 37073 Göttingen, Germany (N.R.).
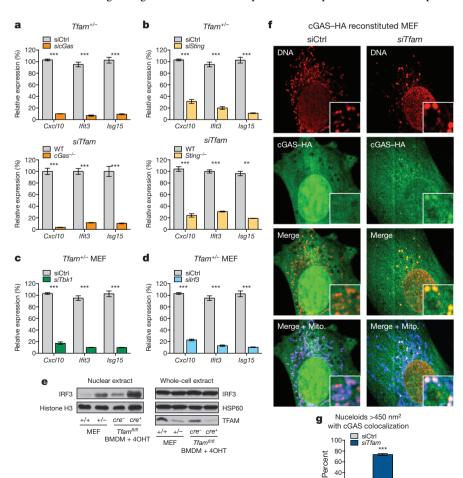
To ensure that the mtDNA stress and ISG expression phenotypes were not unique to $Tfam^{+/-}$ MEFs, we employed inducible TFAM depletion models (TF^D). Analogous to $Tfam^{+/-}$ cells, TF^D MEFs and bone-marrow-derived macrophages (BMDMs) displayed mtDNA stress phenotypes, augmented ISG expression, and heightened type I interferon responses to poly(I:C) (Extended Data Fig. 1d–i). Collectively, these data indicate that TFAM depletion induces mtDNA nucleoid stress that triggers antiviral 'priming', characterized by basally elevated ISG expression and potentiated type I interferon production.

Since mitochondrial stress can trigger the release of mtDNA into the cytosol to engage the NLRP3 inflammasome, we assayed for extra-mitochondrial mtDNA in TF^D cells[9,10]. Analysis of pure cytosolic extracts revealed a three- to fourfold increase of specific mtDNA fragments from the D-loop regulatory region, indicating liberation of immunostimulatory mtDNA into the cytosol (Extended Data Fig. 2)[11]. Confocal and electron microscopy of TF^D cells revealed significantly elongated, interconnected mitochondrial networks consistent with a hyperfused phenotype (Fig. 1a and Extended Data Figs 1e, g and 3a, b). Since mitochondrial fission facilitates proper nucleoid distribution and removal of damaged mtDNA, we examined whether mitochondrial hyperfusion in TF^D cells governed mtDNA stress-induced ISG expression[12,13]. Knockdown of mitofusin 1 (Mfn1) induced fission and largely abrogated ISG expression in TF^D MEFs (Extended Data Fig. 3c–e). Moreover, depletion of the mtDNA quality-control enzyme endo/exonuclease (5′–3′), endonuclease G-like (EXOG) exacerbated ISG expression in $Tfam^{+/-}$ MEFs (Extended Data Fig. 3f)[14]. Collectively, these data indicate that TFAM depletion promotes accumulation of aberrant mtDNA, which accesses the cytosol to engage innate immune signalling.

We next examined the involvement of the cytosolic DNA sensor cGAS in mtDNA stress signalling, as it mediates ISG expression in response to exogenous and endogenous immunostimulatory DNA species[15–17]. Knockdown of cGAS in $Tfam^{+/-}$ MEFs or TFAM depletion in $cGas^{-/-}$ MEFs largely abrogated ISG expression (Fig. 2a). Furthermore, ISG mRNAs in TF^D cells were reduced 70–90% in the absence of STING, indicating cGAS–STING signalling is the predominant driver of mtDNA stress-induced ISG expression (Fig. 2b). STING signals via the TBK1–IRF3/7 axis to trigger antiviral gene expression, and knockdown of either TBK1 or IRF3 robustly blocked ISG expression in $Tfam^{+/-}$ MEFs (Fig. 2c, d)[18,19]. Consistent with IRF3 activating ISG transcription, we observed enhanced nuclear accumulation of IRF3 in TF^D cells (Fig. 2e). Finally, using $cGas^{-/-}$ MEFs reconstituted with hemagglutinin (HA)-tagged, murine cGAS, we observed prominent re-localization of cGAS from nuclear and/or cytoplasmic pools to the vicinity of aberrant mtDNA nucleoids in TF^D MEFs (Fig. 2f, g). Taken together, these results indicate that mtDNA stress facilitates cGAS-dependent sensing of cytoplasmic mtDNA, resulting in STING–TBK1–IRF3 signalling to trigger ISG expression.

To establish functional significance of mtDNA stress-induced antiviral priming, we challenged MEFs with herpes simplex virus 1 (HSV-1) or vesicular stomatitis virus (VSV) that express green fluorescent protein (GFP) for easy detection. In contrast to wild-type cells, which displayed robust viral GFP expression post-infection, $Tfam^{+/-}$ MEFs were markedly resistant to HSV-1 and VSV (Fig. 3a). In addition, $Tfam^{+/-}$ MEFs exhibited heightened type I interferon and ISG expression upon viral challenge, consistent with potentiated type I interferon responses in these cells (Extended Data Fig. 4a). Similar results were obtained upon challenge with the rodent gammaherpesvirus MHV-68 (Fig. 3b and Extended Data Fig. 4b). Furthermore, TF^D BMDMs displayed augmented antiviral gene expression and markedly lower HSV-1- and VSV-encoded mRNA and protein 6–24 h post-infection (Extended Data Fig. 4c–f).

Figure 2 | mtDNA stress triggers ISG expression in a cGAS- and STING-dependent fashion. **a**, **b**, ISG expression in $Tfam^{+/-}$ MEFs transfected with the indicated short interfering RNAs (siRNAs; top panels), or wild-type (WT), $cGas^{-/-}$ (**a**), and $Sting^{-/-}$ (**b**) MEFs transfected with TFAM siRNAs (bottom panels). Ctrl, control. **c**, **d**, ISG expression in $Tfam^{+/-}$ MEFs transfected with the indicated siRNAs for 96 h. **e**, Western blots of whole-cell and nuclear extracts of wild-type and $Tfam^{+/-}$ MEFs or $Tfam^{fl/fl}$ ER-cre$^{+/-}$ (cre$^{+/-}$) BMDMs exposed to 4-hydroxytamoxifen (4OHT) for 96 h. **f**, **g**, $cGas^{-/-}$ MEFs reconstituted with cGAS–HA were transfected with the indicated siRNAs for 96 h, then stained with anti-DNA (DNA), anti-HSP60 (Mito.) and anti-HA (cGAS–HA) antibodies and imaged. cGAS co-localization scoring was performed as described in the Methods. Error bars indicate ±s.e.m. of triplicate technical (**a**–**d**) or biological (**g**) replicates and are representative of three independent experiments. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.
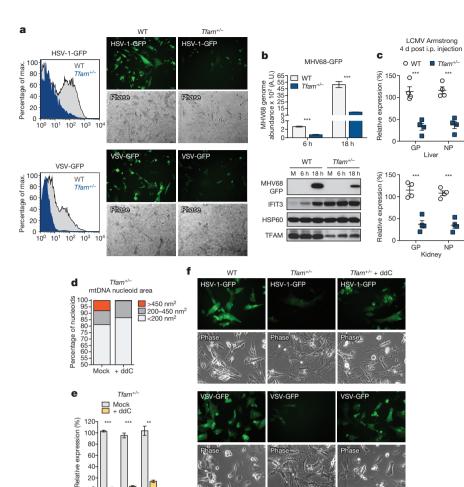
Figure 3 | mtDNA stress potentiates viral resistance. a, Viral GFP expression in MEFs infected with HSV-1-GFP or VSV-GFP at multiplicity of infection (MOI) 0.5 for 24 h. **b**, MHV68-GFP abundance and ISG expression in MEFs infected at MOI 0.5. A.U., arbitrary units. **c**, LCMV Armstrong glycoprotein (GP) and nucleoprotein (NP) gene expression 4 days after intraperitoneal (i.p.) infection of wild-type (WT) and $Tfam^{+/-}$ mice; $n = 4$. **d**, **e**, Nucleoid area (**d**) or ISG expression (**e**) of MEFs exposed to ddC for 96 h. **f**, ddC-exposed MEFs were infected with HSV-1-GFP or VSV-GFP at MOI 0.1 and imaged after 24 h. Error bars represent ±s.e.m. of triplicate technical (**b**, **e**) or quadruplicate biological (**c**) replicates and are representative of two independent experiments. **$P < 0.01$; ***$P < 0.001$.

Finally, we found that $Tfam^{+/-}$ mice exhibit basally elevated ISG expression, which confers resistance to acute infection by lymphocytic choriomeningitis virus (LCMV) Armstrong (Extended Data Fig. 5a and Fig. 3c).

To probe a direct requirement for mtDNA stress in antiviral priming in TFAM-deficient cells, we used dideoxycytidine (ddC), a deoxyribonucleoside analogue that specifically inhibits mtDNA replication and decreases mtDNA nucleoid size[2,20]. Treatment of wild-type MEFs with ddC resulted in reduced mtDNA copy number and decreased average nucleoid size without altering basal ISG expression (Extended Data Fig. 5b–d). In contrast, ddC drastically diminished mtDNA stress (that is, enlarged nucleoids measuring greater than 450 nm^2) in $Tfam^{+/-}$ and TF^D MEFs (Fig. 3d and Extended Data Fig. 5e), which was accompanied by attenuation of antiviral priming and basal ISG expression (Fig. 3e and Extended Data Fig. 5d, f). Moreover, ddC ablated the viral resistance phenotype of $Tfam^{+/-}$ MEFs (Fig. 3f). We observed similar decreases in type I interferon production and a reduction in the viral resistance phenotype in ddC-treated TF^D BMDMs (Extended Data Fig. 5g, h, blue bars). These results demonstrate that mtDNA stress directly potentiates antiviral innate immunity.

The observation that ddC-treated wild-type BMDMs displayed reduced $Ifnb$ and increased viral gene expression upon challenge with HSV-1, despite normal responses to cytosolic nucleic acids (Extended Data Fig. 5h, i, grey bars), indicates that virus-induced mtDNA stress may boost host antiviral responses, consistent with reports linking viral infection to mtDNA dysregulation[21,22]. The alphaherpesvirus protein UL12.5, encoded by HSV-1 and HSV-2, localizes to mitochondria and promotes rapid mtDNA depletion in human cells, which we confirmed in MEFs (Extended Data Fig. 6a)[22–24]. Since mtDNA depletion and nucleoid stress are often coupled, we explored nucleoid architecture and

abundance kinetically during HSV-1 infection. Notably, 3 h after challenge with HSV-1, mtDNA stress was readily apparent, with nucleoids less evenly distributed and enlarged (Fig. 4a). After 6 h, ~10% of nucleoids measured larger than 450 nm^2, and there was a significant decrease in total nucleoid intensity (Fig. 4b). After 12 h, we observed pronounced mtDNA depletion. The mtDNA stress observed 3 to 6 h after HSV-1 challenge closely mirrored that of TFAM-deficient cells (Fig. 4b), as did TFAM protein levels (Fig. 4c). MHV-68 and HSV-2 triggered mtDNA stress similar to HSV-1, indicating that mtDNA stress is a common cellular perturbation during herpesvirus infection (Extended Data Fig. 6b, c). However, induction of mtDNA stress and TFAM depletion were not a general consequence of viral infection, as cells infected with VSV, influenza, LCMV or vaccinia possessed normal mtDNA architecture, TFAM expression and copy number (Fig. 4a–c and Extended Data Fig. 6c, d).

Finally, we sought to determine whether HSV-1-induced mtDNA dysregulation is necessary to fully engage antiviral signalling. Transduction of MEFs and BMDMs with replication-incompetent retroviruses encoding only the mitochondria-targeted HSV-1 UL12 M185 gene product was sufficient to cause mitochondrial hyperfusion, nucleoid enlargement and mtDNA loss, indicative of mtDNA stress (Fig. 4d and Extended Data Fig. 7a)[24]. UL12 M185 expression was also sufficient to trigger TFAM depletion and antiviral priming (that is, augmented ISG mRNA and protein expression) (Fig. 4e and Extended Data Fig. 7a). To explore the effect of HSV-1-induced mtDNA stress on innate antiviral responses, we employed a recombinant, UL12-deficient HSV-1 strain (ΔUL12 + UL98–FLAG) that is severely impaired in its ability to induce mtDNA stress but replicates similarly to a matched UL12-sufficient strain (Extended Data Fig. 7b, c)[25]. Infection with ΔUL12 HSV-1 resulted in attenuated TBK1 phosphorylation and type I interferon

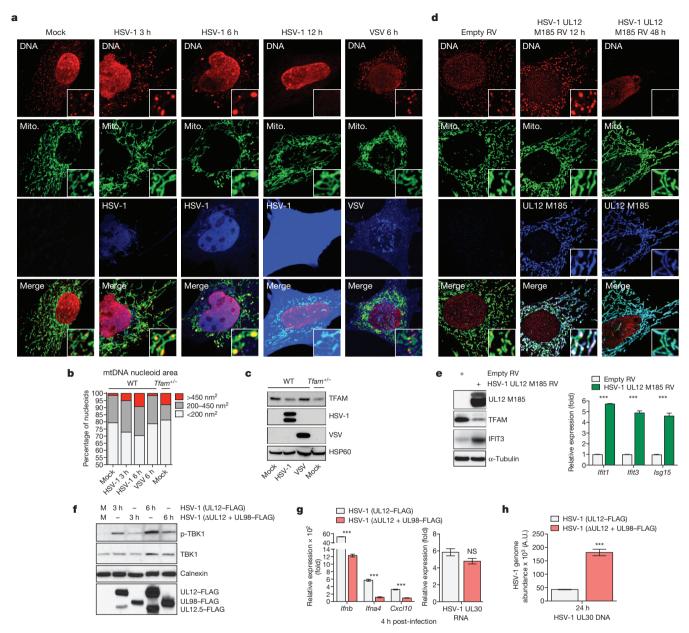**Figure 4 | HSV-1 induces mtDNA stress and TFAM depletion sufficient to trigger ISG expression and necessary to fully engage antiviral immunity.** **a–c**, Wild-type (WT) MEFs were mock infected or infected with HSV-1-GFP or VSV-GFP at multiplicity of infection (MOI) 10 for the indicated times, and imaged after staining with anti-DNA (DNA), anti-HSP60 (Mito.), and anti-HSV (HSV-1) or GFP (VSV) antibodies (**a**). mtDNA nucleoid area was calculated as described in the Methods (**b**). Extracts were blotted as indicated (**c**). **d**, **e**, Wild-type MEFs were transduced with HSV-1 UL12 M185–FLAG-expressing or empty retroviruses (RV) and cells were stained with anti-DNA (DNA), anti-HSP60 (Mito.) and anti-FLAG antibody (UL12 M185) (**d**), and protein or ISG expression examined after 24 h (**e**). **f**, **g**, Protein and RNA expression in BMDMs infected with HSV-1 (UL12–FLAG) or UL12-deficient HSV-1 (ΔUL12 + UL98–FLAG) at MOI 2 for the indicated times. **h**, HSV-1 genome abundance in L929 cells that were infected as in **f**, **g**. A.U., arbitrary units. Error bars indicate ±s.e.m. of triplicate technical replicates and are representative of two independent experiments. ***$P < 0.001$; NS, not significant.

and ISG expression between 3 to 6 h post-infection, despite comparable early HSV-1 gene expression (Fig. 4f, g). However, after 24 h, ΔUL12 HSV-1 genome abundance was roughly threefold higher compared to the UL12-sufficient control, consistent with impaired antiviral innate immunity (Fig. 4h). Finally, ΔUL12 HSV-1 elicited less robust antiviral innate immune responses in the vagina and more readily spread to dorsal root ganglia of wild-type mice due to a deficit in mtDNA stress-dependent antiviral priming (Extended Data Fig. 7d, e). These results reveal that herpesvirus-induced mtDNA stress is necessary to effectively engage ISG expression and antiviral priming, and suggest that cellular monitoring of mtDNA homeostasis represents an additional sensory mechanism to robustly engage antiviral innate immunity.

In closing, our work uncovers a novel cellular response to mtDNA stress that engages the antiviral innate immune response. Specifically, we show that mtDNA stress, induced by herpesvirus infection and mediated by loss of the mtDNA packaging protein TFAM, triggers a cGAS–STING–IRF3-dependent pathway to upregulate ISGs and potentiate type I interferon responses to viral infection (Extended Data Fig. 8). Our results support a model whereby viral-mediated disruption of mtDNA homeostasis serves as a cell-intrinsic indicator of infection that works in parallel with canonical virus sensing to enhance antiviral innate immunity. Conversely, pathologic type I interferon signatures promote autoimmune diseases such as systemic lupus erythematosus, and altered ISG expression correlates with radiation-resistant and

metastatic phenotypes in some cancers[26,27]. Mitochondrial and mtDNA dysregulation have been noted in systemic lupus erythematosus, and perturbations in TFAM and/or mtDNA homeostasis are frequently observed in cancer[28–30]. Therefore, further investigation into this pathway will not only expand our knowledge of innate antiviral defence, but may also broaden our understanding of how mitochondria contribute to the pathogenesis of human diseases and ageing beyond their well characterized roles in metabolism, apoptosis and reactive oxygen species production.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Spelbrink, J. N. Functional organization of mammalian mitochondrial DNA in nucleoids: history, recent developments, and future challenges. *IUBMB Life* **62,** 19–32 (2010).
2. Kasashima, K., Sumitani, M. & Endo, H. Human mitochondrial transcription factor A is required for the segregation of mitochondrial DNA in cultured cells. *Exp. Cell Res.* **317,** 210–220 (2011).
3. Ryan, M. T. & Hoogenraad, N. J. Mitochondrial–nuclear communications. *Annu. Rev. Biochem.* **76,** 701–722 (2007).
4. Wallace, D. C. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.* **39,** 359–407 (2005).
5. Larsson, N. G. et al. Mitochondrial transcription factor A is necessary for mtDNA maintenance and embryogenesis in mice. *Nature Genet.* **18,** 231–236 (1998).
6. Woo, D. K. et al. Mitochondrial genome instability and ROS enhance intestinal tumorigenesis in APC[Min/+] mice. *Am. J. Pathol.* **180,** 24–31 (2012).
7. Rusinova, I. et al. Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41,** D1040–D1046 (2013).
8. Schoggins, J. W. & Rice, C. M. Interferon-stimulated genes and their antiviral effector functions. *Curr. Opin. Virol.* **1,** 519–525 (2011).
9. West, A. P., Shadel, G. S. & Ghosh, S. Mitochondria in innate immune responses. *Nature Rev. Immunol.* **11,** 389–402 (2011).
10. Shimada, K. et al. Oxidized mitochondrial DNA activates the NLRP3 inflammasome during apoptosis. *Immunity* **36,** 401–414 (2012).
11. Nicholls, T. J. & Minczuk, M. In D-loop: 40 years of mitochondrial 7S DNA. *Exp. Gerontol.* **56,** 175–181 (2014).
12. Ban-Ishihara, R., Ishihara, T., Sasaki, N., Mihara, K. & Ishihara, N. Dynamics of nucleoid structure regulated by mitochondrial fission contributes to cristae reformation and release of cytochrome c. *Proc. Natl Acad. Sci. USA* **110,** 11863–11868 (2013).
13. Malena, A., Loro, E., Di Re, M., Holt, I. J. & Vergani, L. Inhibition of mitochondrial fission favours mutant over wild-type mitochondrial DNA. *Hum. Mol. Genet.* **18,** 3407–3416 (2009).
14. Cymerman, I. A., Chung, I., Beckmann, B. M., Bujnicki, J. M. & Meiss, G. EXOG, a novel paralog of Endonuclease G in higher eukaryotes. *Nucleic Acids Res.* **36,** 1369–1379 (2008).
15. Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science* **339,** 786–791 (2013).
16. Ablasser, A. et al. TREX1 deficiency triggers cell-autonomous immunity in a cGAS-dependent manner. *J. Immunol.* **192,** 5993–5997 (2014).
17. Cai, X., Chiu, Y.-H. & Chen, Z. J. The cGAS–cGAMP–STING pathway of cytosolic DNA sensing and signaling. *Mol. Cell* **54,** 289–296 (2014).
18. Atianand, M. K. & Fitzgerald, K. A. Molecular basis of DNA recognition in the immune system. *J. Immunol.* **190,** 1911–1918 (2013).
19. Goubau, D., Deddouche, S. & Reis e Sousa, C. Cytosolic sensing of viruses. *Immunity* **38,** 855–869 (2013).
20. Pohjoismäki, J. L. O. et al. Alterations to the expression level of mitochondrial transcription factor A, TFAM, modify the mode of mitochondrial DNA replication in cultured human cells. *Nucleic Acids Res.* **34,** 5815–5828 (2006).
21. Wiedmer, A. et al. Epstein–Barr virus immediate-early protein Zta co-opts mitochondrial single-stranded DNA binding protein to promote viral and inhibit mitochondrial DNA replication. *J. Virol.* **82,** 4647–4655 (2008).
22. Saffran, H. A., Pare, J. M., Corcoran, J. A., Weller, S. K. & Smiley, J. R. Herpes simplex virus eliminates host mitochondrial DNA. *EMBO Rep.* **8,** 188–193 (2007).
23. Corcoran, J. A., Saffran, H. A., Duguay, B. A. & Smiley, J. R. Herpes simplex virus UL12.5 targets mitochondria through a mitochondrial localization sequence proximal to the N terminus. *J. Virol.* **83,** 2601–2610 (2009).
24. Duguay, B. A. & Smiley, J. R. Mitochondrial nucleases ENDOG and EXOG participate in mitochondrial DNA depletion initiated by herpes simplex virus 1 UL12.5. *J. Virol.* **87,** 11787–11797 (2013).
25. Duguay, B. A. et al. Elimination of mitochondrial DNA is not required for herpes simplex virus 1 replication. *J. Virol.* **88,** 2967–2976 (2014).
26. Crow, M. K. & Kirou, K. A. Interferon-α in systemic lupus erythematosus. *Curr. Opin. Rheumatol.* **16,** 541–547 (2004).
27. Khodarev, N. N. et al. Signal transducer and activator of transcription 1 regulates both cytotoxic and prosurvival functions in tumor cells. *Cancer Res.* **67,** 9214–9220 (2007).
28. Lee, H.-T. et al. Leukocyte mitochondrial DNA alteration in systemic lupus erythematosus and its relevance to the susceptibility to lupus nephritis. *Int. J. Mol. Sci.* **13,** 8853–8868 (2012).
29. Lee, H.-M., Sugino, H., Aoki, C. & Nishimoto, N. Underexpression of mitochondrial-DNA encoded ATP synthesis-related genes and DNA repair genes in systemic lupus erythematosus. *Arthritis Res. Ther.* **13,** R63 (2011).
30. Wallace, D. C. Mitochondria and cancer. *Nature Rev. Cancer* **12,** 685–698 (2012).

Author Contributions A.P.W. designed and performed experiments, analysed data, interpreted results and wrote the paper; W.K.H. provided viral stocks, advice on viral infection protocols, and performed *in vivo* HSV-1 infections; M.S. performed LCMV and influenza infections; M.C.T. aided in experimental design and assisted with viral infections; C.M.P. performed experiments and analysed data; M.B. performed steady-state mitochondrial transcript analysis; N.R. assisted with gene expression array analysis; D.A.M. generated *cGas*[−/−] MEFs; B.A.D. and J.R.S. generated and provided HSV-1 UL12 constructs and HSV-1 ΔUL12 viruses; S.M.K. provided reagents and facilities for LCMV infections and interpreted results; S.M.L. and R.E.M. provided reagents and advice and performed viral infections; A.I. supplied reagents, designed experiments, and interpreted results; G.S.S. designed experiments, interpreted results and wrote the paper.

METHODS

Animal strains. $Tfam^{+/-}$ and $Tfam^{fl/fl}$ mice were previously described and maintained on a C57BL/6 background[6,31]. $Tfam^{fl/fl}$ mice were bred to Estrogen receptor (ER)–Cre transgenic mice from Jackson (stock no. 004682) for inducible, 4OHT-mediated deletion. All animal experiments were conducted in compliance with guidelines established by the Yale University Institutional Animal Care and Use Committee.

Antibodies and reagents. Rabbit anti-mouse TFAM polyclonal anti-sera was previously described[6], rabbit anti-VSV polyclonal anti-sera was a gift from J. Rose at Yale University, mouse anti-Viperin was a gift from P. Cresswell at Yale University, and rabbit-anti IFIT3 was a gift from G. Sen at Cleveland Clinic. The following antibodies were obtained commercially: goat anti-HSP60 (N-20) and rabbit anti-calnexin (H-70) (Santa Cruz Biotechnology); mouse and rabbit anti-FLAG (F1804, F7425) (Sigma); mouse anti-DNA (CBL186) (Millipore); mouse anti-GFP (JL-8) (BD Biosciences); rabbit anti-HSV-1/2 (ab9533) and anti-histone H3 (ab1791) (Abcam); rat anti-HA-FITC (11988506001) (Roche); rabbit anti-NLRX1 (17215-1-AP) (Proteintech); mouse anti-α-tubulin (DM1A) (Neomarkers); mouse anti-GAPDH (6C5) (Ambion); and rabbit anti-DDX58 (D14G6), -IFIH1 (D74E4), -STAT1 (9172), -IRF3 (D83B9), -TBK1 (D1B4) and anti-phospho-TBK1 (D52C2) (Cell Signaling Technology). Mouse IFNα enzyme-linked immunosorbent assay (ELISA) and recombinant mouse IFNβ was from PBL Assay Science, and mouse IL-6 ELISA was from eBioscience. All primer sequences and siRNAs used are found in Extended Data Tables 1 and 2.

Cell culture. Primary wild-type, $Tfam^{+/-}$, $Sting^{-/-}$ and $cGas^{-/-}$ MEFs were generated from E12.5–14.5 embryos, maintained in DMEM (Invitrogen) supplemented with 10% FBS (Atlanta Biological), and sub-cultured no more than five passages before experiments. $Sting^{-/-}$ MEFs were provided by G. Barber at the University of Miami[32]. L929 cells were obtained from ATCC and maintained in DMEM supplemented with 10% FBS. siRNA transfection of MEFs was performed with 25 nM siRNA duplexes (Integrated DNA Technologies) and Lipofectamine RNAiMAX reagent (Invitrogen) according to the manufacturer's instructions. ddC (Sigma) was resuspended in PBS, added to MEFs or BMDMs at a final concentration of 10–20 μM, and replenished every 48 h. BMDMs were generated from bone marrow of 8–12-week-old littermate $Tfam^{fl/fl}$ ER-cre$^-$ and $Tfam^{fl/fl}$ ER-cre$^+$ mice and cultured on Petri plates in DMEM containing 10% FBS plus 30% L929 culture media. To induce Cre-mediated deletion, 1 μM 4OHT dissolved in DMSO (Sigma) was added to BMDM cultures on day 6 and incubated for an additional 2–3 days. Cells were then lifted from plates by incubating in cold PBS containing 1 mM EDTA, replated in fresh media containing 10% L929 conditioned media, and allowed to rest overnight before experimentation (for a total of 72 or 96 h of 4OHT exposure). Transfection of interferon-stimulatory DNA (ISD)[33] and poly(I:C) (Sigma) into the cytosol of BMDMs was performed using Lipofectamine 2000 (Invitrogen). In brief, 1×10^6 BMDMs were seeded in 6-well dishes after 4OHT treatment, and transfected the next day with 4 μg ISD per well or 2.5 μg per well of poly(I:C) complexed at a ratio of 2:1 Lipofectamine 2000 to nucleic acid. Poly(I:C) transfection into the cytosol of MEFs was performed as described previously[34].

Viral stocks and infections. VSV-G-GFP[35], HSV-1-GFP[36], MHV-68-GFP, HSV-2[37], vaccinia virus (strain WR) expressing bacteriophage T7 RNA polymerase[38], influenza A PR8 NS1-GFP[39], HSV-1 (UL12–FLAG) and HSV-1 (UL12Δ + UL98–FLAG)[25] were maintained as described previously[34,40,41]. MEFs or BMDMs were infected at the indicated multiplicity of infection (MOI) in serum-free DMEM for 1 h, washed, and incubated for various times. Cells were then fixed and stained for microscopy, lysed for western blot, solubilized in buffer RLT Plus (Qiagen) for RNA extraction, or prepared for FACS analysis. FACS was performed by first trypsinizing MEFs, followed by labelling with LIVE/DEAD Fixable Far Red stain (Molecular Probes). Cells were then fixed with 4% paraformaldehyde, washed, and analysed on a FACSCalibur flow cytometry machine (BD). FACS plots were first gated on live cells before analysing viral GFP fluorescence. Viral gene expression in BMDMs was determined using qRT–PCR as described below, except that after values were normalized against GAPDH cDNA using the $2^{-\Delta\Delta C_T}$ method, all data points were subtracted by one to centre on zero.

LCMV Armstrong infection of wild-type and $Tfam^{+/-}$ mice was performed as described previously[42]. In brief, 10-week-old female mice were infected with 2×10^5 plaque-forming units of virus intraperitoneally, and 4 days post-infection, mice were euthanized, tissues isolated, and total RNA prepared using RNeasy Plus kits (QIAGEN). After generating complementary DNA, samples were subjected to qPCR analysis as described below using published methods[43,44].

***In vivo* HSV-1 infection, dorsal root ganglia isolation and viral titration.** Six-week old female mice were purchased from Charles River Laboratories and treated with Depo Provera (GE Healthcare) 5 days before infection[45]. The vaginal canals of Depo Provera treated mice were swabbed with a Calginate swab (Fisher) and 10^6 plaque-forming units were delivered via pipette tip into the vagina. One day post-infection, vaginal tissue was isolated for RNA extraction. Infected mice were euthanized at indicated time points and dorsal root ganglia were dissected as previously described[46]. Dorsal root ganglia were homogenized using a motorized pestle and total DNA was isolated using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's instructions. Relative HSV-1 genome abundance was determined using primers specific for nuclear *Tert* and HSV-1 *TK* (thymidine kinase).

cGAS–HA and UL12 M185 cloning and retroviral expression. A plasmid encoding HSV-1 UL12 M185 SPA containing a $3 \times$ FLAG tag at the carboxy-terminus was described previously[24]. This construct, or a plasmid encoding murine cGAS–HA (Invivogen), was sub-cloned into the pMXs-IRES-Puro vector and replication incompetent retroviruses were packaged using plat-E cells according to the manufacturer's instructions (Cell Biolabs). SV40 large T immortalized $cGAS^{-/-}$ MEFs were exposed to supernatants containing cGAS–HA retroviruses and incubated overnight. Two days post-transduction, 3 μg ml^{-1} puromycin was added to select a stable population of cells expressing cGAS–HA. Supernatants containing empty or UL12 M185 SPA retroviruses and 4 μg ml^{-1} polybrene were incubated with cells (5×10^4 MEFs or 2×10^5 BMDMs) in 12-well dishes for a period of 8 h. Viral supernatants were then washed off, fresh media was added to wells, and the cells were incubated for the duration of the experiment until lysis.

Quantitative PCR. To measure relative gene expression by qRT–PCR, total cellular RNA was isolated using RNeasy Plus RNA extraction kit (Qiagen). Approximately 400–2000 ng RNA was normalized across samples and cDNA was generated using the High Capacity cDNA RT kit (Applied Biosystems). cDNA was then subjected to qPCR using Fast SYBR Green Master Mix (Applied Biosystems) and primers as indicated on the ViiA7 Real Time PCR system (Life Technologies). Three technical replicates were performed for each biological sample, and expression values of each replicate were normalized against GAPDH cDNA using the $2^{-\Delta\Delta C_T}$ method. For relative expression (fold), control samples were centred at 1; for relative expression (%), control samples were centred at 100%. Mitochondrial DNA copy number analysis was performed as described using primers specific to nuclear *Tert* and the D-loop region of mtDNA (listed in Extended Data Table 1)[6]. Relative HSV-1 genome abundance was determined using primers specific for nuclear *Tert* and HSV-1 *UL30* or *TK*. Relative MHV68 genome abundance was determined using primers specific for nuclear *Tert* and MHV68 *orf40*. Relative vaccinia genome abundance was determined using primers specific for nuclear *Tert* and vaccinia virus DNA polymerase *E9L*.

Immunofluorescence microscopy. For all microscopy images containing mtDNA nucleoids and associated panels, cells were grown on coverslips and transfected, treated, and/or infected as described. After washing in PBS, cells were fixed with 4% paraformaldehyde for 20 min, permeabilized with 0.1% Triton X-100 in PBS for 5 min, blocked with PBS containing 10% FBS for 30 min, stained with primary antibodies for 60 min, and stained with secondary antibodies for 60 min. Cells were washed with PBS between each step. Coverslips were mounted with Prolong Gold anti-fade reagent containing DAPI (4′,6-diamidino-2-phenylindole; Molecular Probes). Cells were imaged on a Zeiss LSM 510 META with a $63\times$ water-immersed objective. A digital scan zoom of 3.0 was used to enhance magnification. Images were pseudo-coloured and merged using ImageJ software (NIH). For microscopy images in Fig. 3, MEFs were infected as described and fixed with 4% paraformaldehyde for 20 min. Viral GFP fluorescence and phase contrast images were captured using an Olympus IX-71 inverted scope with a $10\times$ (Fig. 3a) or $20\times$ (Fig. 3f) objective. Viral GFP images were pseudo-coloured using ImageJ.

For nucleoid area quantification, approximately 10–15 unique fields of view from 10 distinct confocal images, comprising between 200–400 nucleoids, were captured at random. After incorporating scale information obtained from the LSM Image Browser (Zeiss), images were made binary and the area of each nucleoid was determined using the 'Analyze Particles' feature of ImageJ. Nucleoids were divided into the three size cut-offs: <200 nm^2; 200–450 nm^2; and >450 nm^2, and the percentage of nucleoids falling within each category was plotted. The percentage of nucleoids >450 nm^2 displaying cGAS co-localization was scored by calculating nucleoid area from 5 distinct images of siCtrl- and *siTfam*-transfected cGAS–HA reconstituted MEFs as described above. Nucleoids larger than 450 nm^2 with a substantial cGAS co-localization signal were scored as positive.

Electron microscopy. MEFs grown in Petri dishes and on coverslips for orientation were fixed in 2.5% gluteraldehyde in 0.1 M sodium cacodylate buffer pH 7.4 for 1 h. The cells were rinsed in sodium cacodylate and those in Petri dishes were scraped and spun down in 2% agar. All samples were fixed in 1% osmium tetroxide for 1 h, stained en masse in 2% uranyl acetate in maleate buffer pH 5.2 for a further hour, rinsed and dehydrated in an ethanol series, and infiltrated with resin (Embed812 EMS) and baked over night at 60 °C. Hardened blocks were cut using a Leica UltraCut UCT. 60-nm sections were collected on formver/carbon-coated grids and contrast stained using 2% uranyl acetate and lead citrate. Samples were viewed on an FEI Tencai Biotwin TEM at 80 Kv. Images were taken using Morada CCD and iTEM (Olympus) software.

For mitochondrial perimeter quantification, approximately 10–15 unique electron microscopy images of each genotype were captured at random. After incorporating scale information from iTEM software, the perimeter along the outer membrane of each mitochondrion was traced and quantified using ImageJ. Mitochondria were divided into the three size cutoffs: <2 μm; 2–5 μm; and >5 μm, and the percentage of mitochondria falling within each category was plotted.

Oxygen consumption analysis. Cells were plated in XF96 plates (SeaHorse Biosciences) at 10,000 cells per well and the next day cellular O_2 consumption was determined in a SeaHorse Bioscience XF96 extracellular flux analyser according to the manufacturer's instructions. Cells were maintained at 37 °C in normal growth medium without serum.

Nuclear fractionation and western blotting. Whole-cell extracts were solubilized in SDS lysis buffer (20 mM Tris-HCl, 1% SDS, pH 7.5, containing protease and phosphatase inhibitors), boiled for 5 min, and DNA was sheared by sonication. For nuclear extraction, PBS-washed cell pellets were resuspended in 10 pellet volumes of RSB buffer (10 mM NaCl, 1.5 mM $CaCl_2$, 10 mM Tris-HCl pH 7.5), swelled on ice for 10 min, homogenized with a motorized Teflon pestle, and the homogenate was centrifuged at 980g for 10 min to pellet nuclei. Pellets were washed five times in PBS, SDS was then added to a final concentration of 1%, and extracts were boiled for 5 min before sonicating to shear DNA and normalizing protein concentration. Western blotting was performed using standard protocols, and HSP60 was used as whole-cell extract loading controls, while histone H3 was probed as a nuclear loading control.

Detection of mtDNA in cytosolic extracts. Digitonin extracts from MEFs and BMDMs were generated largely as described previously[47]. Wild-type and $Tfam^{+/-}$ MEFs (7×10^6) or $Tfam^{fl/fl}$ ER-$cre^{+/-}$ BMDMs exposed to 4OHT for 72 h (1×10^7) were each divided into two equal aliquots, and one aliquot was resuspended in 500 μl of 50 μM NaOH and boiled for 30 min to solubilize DNA. 50 μl of 1 M Tris-HCl pH 8 was added to neutralize the pH, and these extracts served as normalization controls for total mtDNA. The second equal aliquots were resuspended in roughly 500 μl buffer containing 150 mM NaCl, 50 mM HEPES pH 7.4, and 15–25 μg ml^{-1} digitonin (EMD Chemicals). The homogenates were incubated end over end for 10 min to allow selective plasma membrane permeabilization, then centrifuged at 980g for 3 min three times to pellet intact cells. The first pellet was saved as the 'Pel' fraction for western blotting. The cytosolic supernatants were transferred to fresh tubes and spun at 17000g for 10 min to pellet any remaining cellular debris, yielding cytosolic preparations free of nuclear, mitochondrial and endoplasmic reticulum contamination. DNA was then isolated from these pure cytosolic fractions using QIAQuick Nucleotide Removal Columns (QIAGEN). Quantitative PCR was performed on both whole-cell extracts and cytosolic fractions using nuclear DNA primers (*Tert*) and mtDNA primers (*Dloop1-3*, *Cytb*, *16S* and *Nd4*), and the C_T values obtained for mtDNA abundance for whole-cell extracts served as normalization controls for the mtDNA values obtained from the cytosolic fractions. This allowed effective standardization among samples and controlled for any variations in the total amount of mtDNA in control and TFAM-deficient samples. Using this digitonin method, no nuclear *Tert* DNA was detected in the cytosolic fractions, indicating nuclear lysis did not occur.

Bioinformatic analyses. Total cellular RNA from wild-type and $Tfam^{+/-}$ littermate MEFs was prepared using RNeasy Plus RNA extraction kits (QIAGEN) and used for the expression microarray procedure in conjunction with the Emory University Integrated Genomics Core. RNA integrity was first verified by an Agilent Bioanalyzer and then amplified, labelled, and hybridized onto Mouse Gene 1.0 ST arrays (Affymetrix) using standard protocols recommended by the manufacturer, starting from 2 μg of total RNA. Data were normalized by the RMA method using GeneSpring software (Agilent) for each biological sample in duplicate. Student's t-test was used to determine statistically significant changes in expression in $Tfam^{+/-}$ MEFs relative to wild type, with a cut-off P value of 0.05[48]. Heat maps were generated using MultiExperiment Viewer[49].

Statistical analyses. Error bars displayed throughout the manuscript represent s.e.m. unless otherwise indicated, and were calculated from triplicate or quadruplicate technical replicates of each biological sample. For *in vivo* experiments, error bars were calculated from the average of triplicate technical replicates of 3–4 mice per point. Sample sizes were chosen by standard methods to ensure adequate power, and no randomization or blinding was used for animal studies. No statistical method was used to predetermine sample size. Statistical significance was determined using unpaired Student's t-tests; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; NS, not significant ($P > 0.05$). Data shown are representative of 2–3 independent experiments, including microscopy images, western blots and viral challenges.

31. Weinberg, F. *et al.* Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity. *Proc. Natl Acad. Sci. USA* **107,** 8788–8793 (2010).
32. Ishikawa, H. & Barber, G. N. STING is an endoplasmic reticulum adaptor that facilitates innate immune signalling. *Nature* **455,** 674–678 (2008).
33. Stetson, D. B. & Medzhitov, R. Recognition of cytosolic DNA activates an IRF3-dependent innate immune response. *Immunity* **24,** 93–103 (2006).
34. Tal, M. C. *et al.* Absence of autophagy results in reactive oxygen species-dependent amplification of RLR signaling. *Proc. Natl Acad. Sci. USA* **106,** 2770–2775 (2009).
35. Dalton, K. P. & Rose, J. K. Vesicular stomatitis virus glycoprotein containing the entire green fluorescent protein on its cytoplasmic domain is incorporated efficiently into virus particles. *Virology* **279,** 414–421 (2001).
36. Desai, P. & Person, S. Incorporation of the green fluorescent protein into the herpes simplex virus type 1 capsid. *J. Virol.* **72,** 7563–7568 (1998).
37. Shin, H. & Iwasaki, A. A vaccine strategy that protects against genital herpes by establishing local memory T cells. *Nature* **491,** 463–467 (2012).
38. Fuerst, T. R., Niles, E. G., Studier, F. W. & Moss, B. Eukaryotic transient-expression system based on recombinant vaccinia virus that synthesizes bacteriophage T7 RNA polymerase. *Proc. Natl Acad. Sci. USA* **83,** 8122–8126 (1986).
39. Pang, I. K., Pillai, P. S. & Iwasaki, A. Efficient influenza A virus replication in the respiratory tract requires signals from TLR7 and RIG-I. *Proc. Natl Acad. Sci. USA* **110,** 13910–13915 (2013).
40. Yordy, B., Iijima, N., Huttner, A., Leib, D. & Iwasaki, A. A neuron-specific role for autophagy in antiviral defense against herpes simplex virus. *Cell Host Microbe* **12,** 334–345 (2012).
41. Cardenas, I. *et al.* Placental viral infection sensitizes to endotoxin-induced pre-term labor: a double hit hypothesis. *Am. J. Reprod. Immunol.* **65,** 110–117 (2011).
42. Marshall, H. D. *et al.* Differential expression of Ly6C and T-bet distinguish effector and memory Th1 CD4$^+$ cell properties during viral infection. *Immunity* **35,** 633–646 (2011).
43. Welsh, R. M. & Seedhom, M. O. Lymphocytic choriomeningitis virus (LCMV): propagation, quantitation, and storage. *Curr. Protoc. Microbiol.* **Unit 15A. 1,** http://dx.doi.org/10.1002/9780471729259.mc15a01s8 (2008).
44. McCausland, M. M. & Crotty, S. Quantitative PCR technique for detecting lymphocytic choriomeningitis virus in vivo. *J. Virol. Methods* **147,** 167–176 (2008).
45. Parr, M. B. *et al.* A mouse model for studies of mucosal immunity to vaginal infection by herpes simplex virus type 2. *Lab. Invest.* **70,** 369–380 (1994).
46. Malin, S. A., Davis, B. M. & Molliver, D. C. Production of dissociated sensory neuron cultures and considerations for their use in studying neuronal function and plasticity. *Nature Protocols* **2,** 152–160 (2007).
47. Holden, P. & Horton, W. A. Crude subcellular fractionation of cultured mammalian cell lines. *BMC Res. Notes* **2,** 243 (2009).
48. Raimundo, N. *et al.* Mitochondrial stress engages E2F1 apoptotic signaling to cause deafness. *Cell* **148,** 716–726 (2012).
49. Saeed, A. I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34,** 374–378 (2003).

Super-enhancers delineate disease-associated regulatory nodes in T cells

Golnaz Vahedi[1], Yuka Kanno[1], Yasuko Furumoto[2], Kan Jiang[1], Stephen C. J. Parker[3]†, Michael R. Erdos[3], Sean R. Davis[4], Rahul Roychoudhuri[4], Nicholas P. Restifo[4], Massimo Gadina[2], Zhonghui Tang[5], Yijun Ruan[5], Francis S. Collins[3], Vittorio Sartorelli[6] & John J. O'Shea[1]

Enhancers regulate spatiotemporal gene expression and impart cell-specific transcriptional outputs that drive cell identity[1]. Super-enhancers (SEs), also known as stretch-enhancers, are a subset of enhancers especially important for genes associated with cell identity and genetic risk of disease[2–6]. CD4[+] T cells are critical for host defence and autoimmunity. Here we analysed maps of mouse T-cell SEs as a non-biased means of identifying key regulatory nodes involved in cell specification. We found that cytokines and cytokine receptors were the dominant class of genes exhibiting SE architecture in T cells. Nonetheless, the locus encoding *Bach2*, a key negative regulator of effector differentiation, emerged as the most prominent T-cell SE, revealing a network in which SE-associated genes critical for T-cell biology are repressed by BACH2. Disease-associated single-nucleotide polymorphisms for immune-mediated disorders, including rheumatoid arthritis, were highly enriched for T-cell SEs versus typical enhancers or SEs in other cell lineages[7]. Intriguingly, treatment of T cells with the Janus kinase (JAK) inhibitor tofacitinib disproportionately altered the expression of rheumatoid arthritis risk genes with SE structures. Together, these results indicate that genes with SE architecture in T cells encompass a variety of cytokines and cytokine receptors but are controlled by a 'guardian' transcription factor, itself endowed with an SE. Thus, enumeration of SEs allows the unbiased determination of key regulatory nodes in T cells, which are preferentially modulated by pharmacological intervention.

Histone acetyltransferase p300 loading demarcates regions of the genome bearing SE architecture[2,8]. Using chromatin immunoprecipitation followed by sequencing (ChIP-seq) for the p300 protein, we constructed SE catalogues of murine CD4[+] T helper (T_H)1, T_H2 and T_H17 cells. As predicted[2], the p300 load is exponentially distributed throughout the genome (Fig. 1a and Extended Data Fig. 1a). Approximately 40% of the p300 signal was found in a small fraction of p300-loaded enhancers in each lineage. The distribution of SEs was lineage-specific even in these closely related cells (Fig. 1b and Extended Data Fig. 1b). Regulatory regions of lineage-specific master transcription factors were endowed with SEs only in the relevant lineage (Extended Data Fig. 1c). We addressed the relationship between SEs and transcriptional activity in T cells by assigning SEs to associated genes using proximity measures[4], bearing in mind that alternative methods can conclusively establish such associations[6,9]. We found that SE architecture conferred significantly higher transcriptional activity compared with typical enhancer (TE) architecture and that this transcriptional activity was lineage-specific (Fig. 1c, d).

Widespread transcription at SEs themselves has been reported in embryonic stem (ES) cells and myogenic cells[2,10]. We next explored the extent to which SE domains were transcribed in T cells by employing high-resolution temporal expression maps of intergenic noncoding RNAs (ncRNAs)[11]. One-third of the ncRNAs expressed in T cells (501/1,524) were transcribed from an SE[10] (Fig. 1e and Extended Data Fig. 1d).
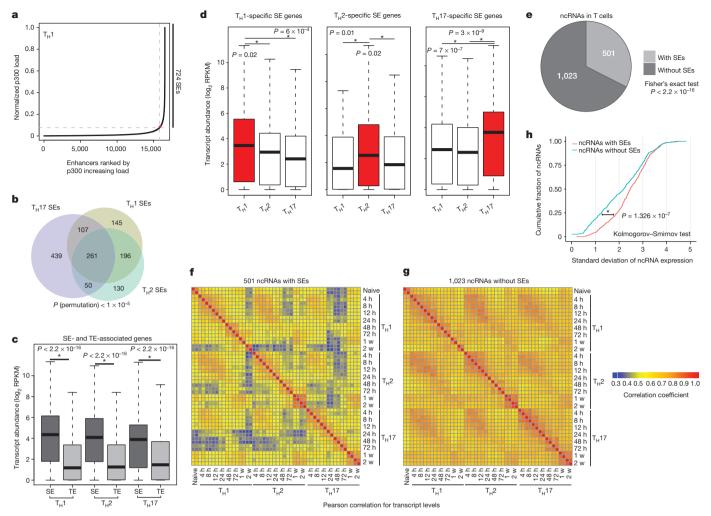
Controlling for differences in the size of SEs and TEs, we found 80 ncRNAs per 10 megabase pairs (Mb) of SEs compared with 51 transcripts within TEs. The presence of an SE structure also distinguished highly lineage-specific and dynamic ncRNAs from constitutively expressed ones (Fig. 1f–h).

To elucidate the potential role of SEs in T-cell biology, we used ChIP-seq data sets to catalogue binding profiles of 13 transcription factors with major roles in T_H-cell differentiation across the merged map of SEs[12–15] (Fig. 2a–c). As in ES cells[2], STATs prominently bound SEs in CD4[+] T cells (Fig. 2a, d). Similarly, BATF, IRF4 and BACH2 were enriched at these regions (Fig. 2b, d). Lineage-specific transcription factors such as T-BET, GATA3 and ROR-γt showed preferential binding at lineage-specific SEs (Extended Data Fig. 2a). CTCF, an essential genome organizer, appeared to preferentially demarcate SE boundaries[6] (Extended Data Fig. 2b). Comparison of the enrichment of transcription factors at SEs and TEs revealed selective binding of STAT3 at SEs whereas other transcription factors demonstrated comparable binding at SEs and TEs (Extended Data Fig. 2c).

We next compared the identity of SE-associated genes in T cells with those in other cell lineages. In ES cells, SE structures primarily encompass transcription factors (Fig. 2e and Extended Data Fig. 3a). In macrophages, chemokine and cytokine activity were the most prominent categories. In T lymphocytes, genes relevant to cytokine biology were preferentially linked to SEs. Moreover, cytokine-related genes were not linked to SEs in non-immune related cells such as myotubes (Extended Data Fig. 3b). Thus, SEs are preferentially associated with genes that have a central role in the biology of specific cell lineages rather than a given class of genes (that is, transcription factors). In the case of T cells, SEs form an interactive network that reflects the biology of lymphocytes, their products and their mode of sensing the inflammatory environment.

We next ranked T-cell SEs on the basis of their p300 occupancy (Fig. 3a). Again, SEs with the highest p300 occupancy were typically associated with genes encoding cytokines and their receptors. However, the greatest p300 enrichment was associated with the *Bach2* locus, regardless of lineage subset (Fig. 3a, b). This is of interest as BACH2 is a broad regulator of immune activation that acts by stabilizing immunoregulatory capacity and attenuating effector differentiation[13]. Notably, genetic variations within this locus are associated with numerous immune-mediated diseases including rheumatoid arthritis[16], Crohn's disease[17], multiple sclerosis[18], asthma[19] and type 1 diabetes[20]. These observations prompted us to investigate the effect of *Bach2* deletion on the expression of SE-associated genes in T cells. Transcriptional profiling revealed that *Bach2* deficiency significantly affected the expression of genes with SE architecture compared to those with TEs or no enhancer mark in T cells (Fig. 3c, d). These findings were confirmed when we used synthetic RNA standards 'spiked-in' to rigorously normalize transcriptome data in wild-type and *Bach2*-deficient cells[21] (Methods; Extended Data

[1]Lymphocyte Cell Biology Section, National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), National Institutes of Health (NIH), Bethesda, Maryland 20892, USA. [2]Translational Immunology Section, NIAMS, NIH, Bethesda, Maryland 20892, USA. [3]Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892, USA. [4]Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA. [5]The Jackson Laboratory for Genomic Medicine and Department of Genetic and Development Biology, University of Connecticut, Farmington, Connecticut 06030, USA. [6]Laboratory of Muscle Stem Cells and Gene Regulation, NIAMS, NIH, Bethesda, Maryland 20892, USA. †Present address: Departments of Computational Medicine & Bioinformatics and Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA.

Figure 1 | SE structure predicts lineage- and stage-specific transcription.
a, The histone acetyltransferase p300 is distributed asymmetrically across the genome in CD4$^+$ T cells with a subset of enhancers (SEs) that bind exceptionally high amounts of p300 (see Source Data). **b**, Closely related CD4$^+$ T-cell populations have distinctive SE landscapes. Venn diagram depicts shared and unique SE domains in T-cell subsets. **c**, SE-associated genes are highly transcribed compared with TE-associated genes. Proximity measures were used to assign SEs and TEs to their target genes (P values, Wilcoxon rank-sum test). RPKM, reads per kilobase of exon per million. **d**, Presence of lineage-specific SEs predicts cell-selective expression. Three groups of genes associated with unique SE structure in each lineage were defined as T_H1-, T_H2- and T_H17-specific SE genes. The expression of lineage-specific SE-associated genes was more significant in the corresponding lineage (P values, Wilcoxon rank-sum test). **e**, SE domains are themselves transcribed in CD4$^+$ T cells. The list of ncRNAs was derived from the map of intergenic transcripts in T-cell subsets[11]. One-third of ncRNAs in T cells (501/1,524) were transcribed from an SE. **f–h**, The SE structure differentiates highly lineage-specific and dynamic noncoding transcripts from constitutively expressed transcripts across T-cell lineages. **f**, **g**, Pearson correlation coefficients for transcription levels between each pair of differentiation stages were calculated for 501 ncRNAs with SEs (**f**) and 1,023 ncRNAs without SEs (**g**). **h**, ncRNA transcripts with SEs have a greater standard deviation across differentiation stages compared to those without SEs.

Fig. 3c, d). This transcriptional difference remained statistically significant when we controlled for higher levels of gene expression for SE-associated genes (Extended Data Fig. 3e). Furthermore, loss of BACH2 led to the largest difference between SEs and TEs in comparison with other transcription factors such as STATs, BATF and IRF4 (Extended Data Fig. 4a, b). In particular, 348 genes, 26% of those with SE structure in CD4$^+$ T cells, were repressed by BACH2 (Fig. 3e and Extended Data Fig. 4c–e). In addition to protein-coding genes, a subset of SE-linked ncRNAs (56) were also repressed by BACH2 (Fig. 3f). Transcriptional upregulation at some of these domains correlated with the upregulation of nearby genes in *Bach2*-deficient cells (Fig. 3g and Extended Data Fig. 4f, g). This previously unrecognized circuitry reveals that a subset of genes and noncoding transcripts endowed with SE architecture in CD4$^+$ T cells are tightly and negatively controlled by the 'guardian' transcription factor BACH2, which itself has a rich cassette of regulatory elements (Extended Data Fig. 4h).

It has been shown that single-nucleotide polymorphisms (SNPs) associated with diseases relevant to a particular cell type are more enriched

in SEs compared with TEs[2,5]. CD4$^+$ T cells are important contributors to a wide variety of autoimmune diseases including rheumatoid arthritis. Thus, we explored the extent to which rheumatoid-arthritis-associated genetic variants were situated within SEs. We delineated SEs in human CD4$^+$ T-cell subsets and found that 26% of the SNPs highly associated with rheumatoid arthritis[7] (27/101) fell within SEs (Fig. 4a). In contrast, only 7% of rheumatoid arthritis SNPs overlapped with TEs (Fig. 4a). Controlling for differences in the size of genomic regions, we found that the number of SNPs per 10 Mb of SEs was significantly higher than that in TEs (Fig. 4a). Genetic variants associated with other autoimmune disorders such as inflammatory bowel disease, multiple sclerosis and type 1 diabetes also exhibited preferential enrichment in CD4$^+$ T-cell SEs compared to TEs (Fig. 4a). Such enrichment was also present when we considered variants in high linkage disequilibrium with disease-associated SNPs (Extended Data Fig. 5a). As a comparison, genetic variants associated with type 2 diabetes and cancer, diseases in which CD4$^+$ T cells are not thought to have major roles, were also assessed and found not to be significantly enriched within T-cell SEs (Fig. 4a).
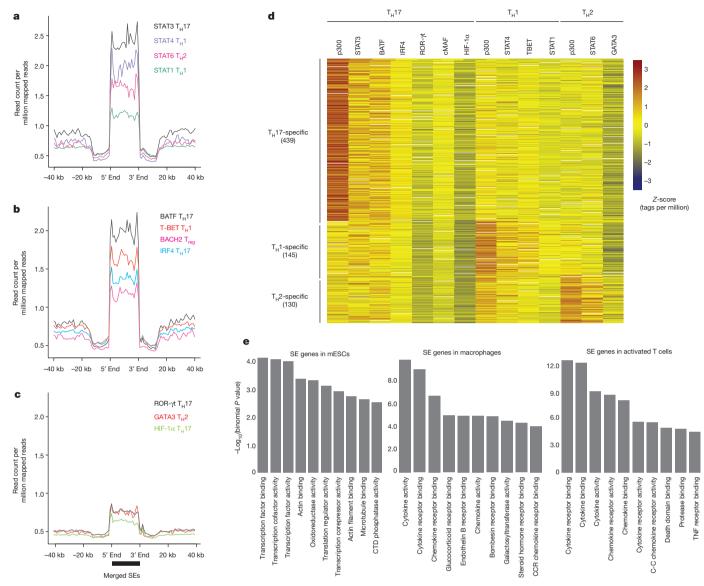
Figure 2 | Transcription factors with major roles in T_H-cell differentiation occupy SEs. a–c, Lineage-predicting transcription factors are enriched at SE domains. The catalogue of SEs in $CD4^+$ T cells was constructed by merging T_H1, T_H2 and T_H17 SEs. Binding patterns of STAT1, STAT3, STAT4 and STAT6 (**a**), BATF, T-BET, BACH2 and IRF4 (**b**), and HIF-1α, ROR-γt and GAT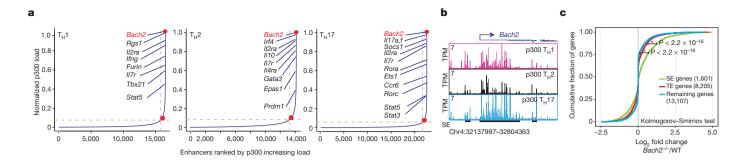A3 (**c**) are demonstrated at SEs. **d,** Binding of lineage-specific transcription factors correlates with the presence of lineage-specific SEs in T cells ($\log_2$ tags per million) (see Source Data). **e,** Gene ontology (GO) functional categories relevant to cytokines and cytokine receptors are enriched at SE-associated genes in T cells. GO analysis for SE regions was performed using GREAT[24]. mESCs, mouse ES cells.

We refined these observations by examining genes that were affected by rheumatoid-arthritis-associated genetic variants, focusing on 98 candidate genes associated with rheumatoid arthritis[7]. While SEs in muscle cells showed little association (Fig. 4b), rheumatoid arthritis risk genes were preferentially associated with SEs in cytotoxic natural killer cells ($CD56^+$) and monocytes ($CD14^+$). However, the strongest enrichment occurred in $CD4^+$ T cells, where half of the rheumatoid arthritis risk genes (53/98) were linked to $CD4^+$ T-cell SEs (Fig. 4b).

SE structures are thought to be particularly sensitive to perturbation owing to the cooperative and synergistic binding of numerous factors at these domains[3]. Given the enrichment of STATs at SEs and the prevalence of SEs at cytokines and their receptors, we measured the effect of tofacitinib, a JAK inhibitor recently approved by the US Food and Drug Administration for the treatment of rheumatoid arthritis, on T-cell transcriptomes. We found that tofacitinib treatment had a significantly greater impact on the transcription of genes with SEs than TEs (Extended Data Fig. 5b). Moreover, when genes were ranked on the basis of their transcript levels in T cells, the most highly expressed genes with SEs

showed a larger change in their expression compared to those without SEs, emphasizing that tofacitinib discriminates genes with SE structure (Extended Data Fig. 5c). Although harbouring the strongest SE in T cells, BACH2 levels were not affected by acute tofacitinib treatment, suggesting that BACH2 is regulated in a JAK/STAT-independent manner. Finally, we related the effect of this rheumatoid arthritis drug to the genetics of the disease and found that tofacitinib treatment disproportionately affected the expression of rheumatoid arthritis risk genes with SE structures in $CD4^+$ T cells compared with those lacking this chromatin feature (Fig. 4c and Extended Data Fig. 5d). Furthermore, tofacitinib treatment selectively affected inflammatory bowel disease[22] and multiple sclerosis[23] risk genes with SEs (Extended Data Fig. 6).

We have defined the T_H-cell SE landscape in the hope of better defining key regulatory nodes in a non-biased fashion. We found that in T cells these nodes largely comprise cytokine and cytokine receptor genes. Thus, T-cell 'identity' relates largely to the precise regulation of these key effectors and sensors. However, a predominant SE-associated gene in all T-cell lineages was *Bach2*, which may represent the first example
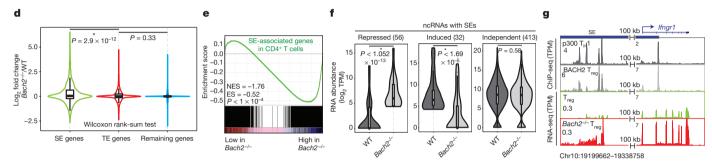
Figure 3 | Bach2 is endowed with the highest p300-enriched SE in T cells. a, Ranked order of p300-loaded enhancers in T-cell subsets identifies *Bach2* as the strongest SE-associated gene in CD4[+] T cells. b, The *Bach2* locus, the top ranked SE, exhibits an exceptional amount of p300 binding. c, d, BACH2 preferentially represses SE genes. Wild-type (WT) and *Bach2*-deficient CD4[+] T cells were polarized to induced regulatory T cells (iT_reg cells) and were processed for total RNA extraction (*n* = 3). Normalized transcript abundance measured by RNA-seq (RPKM) was evaluated in wild-type and *Bach2*-deficient cells at SE- and TE-associated genes and compared to the remaining genes. Cumulative distribution (c) and violin plots (d) show the (log_2) fold change in gene expression for wild-type versus *Bach2*-deficient cells (see Source Data).

e, Gene set enrichment analysis (GSEA) of SE-associated genes reveals that SE genes are enriched in genes repressed by BACH2. ES, enrichment score; NES, normalized enrichment score. f, BACH2 affects a subset of noncoding transcripts at SE domains. Overall, 56 ncRNAs with SE structures are repressed while 32 transcripts are induced by BACH2 (see Source Data). *P* values, Wilcoxon rank-sum test. g, BACH2-associated repression of a noncoding transcript with an SE architecture correlates with the transcriptional repression of a nearby gene (*Ifngr1*). Direct BACH2 binding along with the transcript levels in wild-type and *Bach2*-deficient cells measured by RNA-seq were depicted in a 140 kb window accommodating the *Ifngr1* gene. TPM, tags per million.
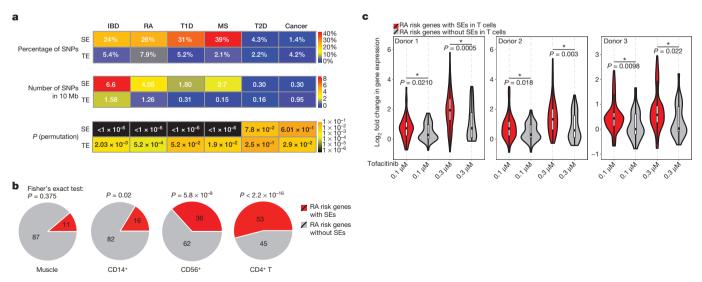

Figure 4 | Rheumatoid arthritis risk genes with SE structure are selectively targeted by the JAK inhibitor tofacitinib. a, SNPs associated with autoimmune diseases including rheumatoid arthritis (RA), inflammatory bowel disease (IBD), multiple sclerosis (MS), and type 1 diabetes (T1D) are preferentially enriched at the SE structure of human CD4[+] T cells. In contrast, SNPs associated with disorders in which CD4[+] T cells have limited roles, such as type 2 diabetes (T2D) and cancer, are not enriched in these genomic domains. A catalogue of 1,426 SEs in human T cells was constructed by aggregating SE predictions in human T_H1, T_H2 and T_H17 cells using H3K27ac data (see Source Data). We divided the number of SNPs enriched in SEs/TEs by the total size of SEs (66.5338 Mb) and TEs (63.12915 Mb) and reported

the number of SNPs within every 10 Mb of the genome (*P* values, permutations test). b, Rheumatoid arthritis risk genes are linked to SEs in CD4[+] T cells. The 98 candidate genes associated with rheumatoid arthritis were from ref. 7. c, Rheumatoid arthritis risk genes with SEs are selectively targeted by a JAK inhibitor, tofacitinib. Violin plots depict the fold change in expression (log_2) after tofacitinib treatment of human CD4[+] T cells at rheumatoid arthritis risk genes with or without SEs (three donors). To ensure accurate inference of the effect of tofacitinib on the transcriptome, spiked-in RNA standards were added and gene expression levels (RPKM) were renormalized based on the spiked-in standards (*P* values, Wilcoxon rank-sum test).

of a class of transcriptional regulators that broadly constrains transcription at SEs. Furthermore, SNPs associated with immune-related diseases were enriched at T-cell SEs, and a drug, which blocks cytokine signalling and is clinically efficacious in autoimmune disease, preferentially impacted SE-associated genes. Hence, our study provides a systematic approach by which the SE map of relevant cell types can be integrated with human genetics to discover drug target genes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
2. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
3. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153,** 320–334 (2013).
4. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).
5. Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110,** 17921–17926 (2013).
6. Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in Mammalian chromosomes. *Cell* **159,** 374–387 (2014).
7. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).
8. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
9. Kieffer-Kwon, K. R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155,** 1507–1520 (2013).
10. Mousavi, K. *et al.* eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell* **51,** 606–617 (2013).
11. Hu, G. *et al.* Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nature Immunol.* **14,** 1190–1198 (2013).
12. Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151,** 289–303 (2012).
13. Roychoudhuri, R. *et al.* BACH2 represses effector programs to stabilize T$_{reg}$-mediated immune homeostasis. *Nature* **498,** 506–510 (2013).
14. Wei, L. *et al.* Discrete roles of STAT4 and STAT6 transcription factors in tuning epigenetic modifications and transcription during T helper cell differentiation. *Immunity* **32,** 840–851 (2010).
15. Vahedi, G. *et al.* STATs shape the active enhancer landscape of T cell populations. *Cell* **151,** 981–993 (2012).
16. McAllister, K. *et al.* Identification of BACH2 and RAD51B as rheumatoid arthritis susceptibility loci in a meta-analysis of genome-wide data. *Arthritis Rheum.* **65,** 3058–3062 (2013).
17. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42,** 1118–1125 (2010).
18. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).
19. Ferreira, M. A. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378,** 1006–1014 (2011).
20. Cooper, J. D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genet.* **40,** 1399–1401 (2008).
21. Lovén, J. *et al.* Revisiting global gene expression analysis. *Cell* **151,** 476–482 (2012).
22. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
23. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genet.* **45,** 1353–1360 (2013).
24. McLean, C. Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotechnol.* **28,** 495–501 (2010).

Author Contributions G.V., V.S., F.S.C. and J.J.O'S. participated in the study design. Y.K., Y.F. and K.J. performed sequencing experiments. Z.T. and Y.R. supervised and performed sequencing experiments. Y.F. and M.G. performed tofacitinib-related experiments. G.V. performed computational analysis. S.C.J.P., M.R.E. and S.R.D. participated in statistical analysis relevant to human genetics. R.R. and N.P.R. supervised and performed experiments involving *Bach2*-deficient cells. Y.K., Y.F. and S.R.D. participated in writing of the methodology. G.V., V.S. and J.J.O'S. wrote the manuscript and all authors reviewed it. J.J.O'S. supervised the project.

METHODS

Antibodies and reagents. The following antibodies and reagents were obtained from eBioscience: CD4-PerCPCy5.5, CD45RA-PE, CD45RO-eFluor450, CD28-purified. Anti-CD3 antibody was obtained from BioXcell. CP-690,550 (tofacitinib) was prepared by the National Institutes of Health (NIH) Chemical Genomics Center and dissolved in dimethylsulphoxide (DMSO). Source Data associated with Fig. 1 summarizes ChIP-seq data sets generated or used for this data set along with relevant antibodies.

Cell culture and stimulation for tofacitinib-treated human T cells. Whole blood from healthy donors was provided from the NIH blood bank and informed consent was obtained from subjects. To obtain lymphocyte population, heparinized whole blood from healthy donors was separated by Ficoll Paque PLUS (Sigma). Naive CD4$^+$ CD45RA$^+$ CD45RO$^-$ T-cell population was sorted on a FACS Aria III (BD Bioscience). Cells were activated by plate-bound anti-CD3/anti-CD28 (10 μg ml^{-1}) in supplemented RPMI 1640 medium containing 10% FCS, 2 mM glutamine, 100 IU ml^{-1} penicillin, 0.1 mg ml^{-1} streptomycin, 20 mM HEPES buffer (all from Invitrogen) for 3 days and cultured in the presence of IL-2 for 1 day. During T-cell activation, cells were treated with the indicated concentrations of CP-690,550 (tofacitinib).

RNA-seq preparation. Total RNA was prepared from approximately 1 million cells by using mirVana miRNA Isolation Kit (AM1560, ABI). Two-hundred nanograms to 1 μg of total RNA was subsequently used to prepare RNA-seq libraries by using TruSeq SR RNA sample prep kit (FC-122-1001, Illumina) or by a combination of NEBNext RNA library prep kit (New England BioLabs) and Ovation SP Ultralow DR Multiplex system (Nugen) by following the manufacturer's protocol. The libraries were sequenced for 50 cycles (single read) with HiSeq 2000 (Illumina). Where indicated, ERCC RNA spike-in mix 1 (Invitrogen) was added to the samples based on the cell counts (1 μl of 1/10 dilution to 1 million cells).

RNA-seq analysis. RNA-seq libraries made by Illumina TruSeq were first trimmed using 'cutadapt' with TruSeq Indexed Adapters. An error rate of 0.1 was chosen for 'cutadapt'. Overall, the percentage of trimmed reads was lower than 3% of the total reads across different libraries. Trimmed fastq files were then aligned to mm9 or hg19 reference genomes using tophat with bowtie2 indexes derived based on UCSC annotations. The normalization of RNA-seq libraries shown on the genome browser was carried out using 'bedtools genomecov' to 'scale' the bam files to tags-per-million values. 'HT-seq' was used to find the read counts across the UCSC reference genome and DEseq was further employed to characterize differentially regulated genes where repeats were available (*Bach2*-deficient RNA-seq).

RNA-seq analysis of *Bach2*-deficient cells. Wild-type and *Bach2*-deficient naive (CD44$^-$ CD62L$^+$ CD25$^-$) CD4$^+$ cells were isolated to >95% purity from C57BL/6 mice reconstituted with mixtures of wild-type and knockout OT-II TCR-transgenic bone marrow. Cells were stimulated at 1×10^5 cells per 96-well plate coated in 5 μg ml^{-1} anti-CD3 in the presence of soluble anti-CD28 (5 μg ml^{-1}), 100 IU recombinant human IL-2 and 5 ng ml^{-1} recombinant human TGF-β for 3 days. Cells were counted using a haematocytometer, or analysed by FACS for cell size or intracellular Foxp3 content. Cells were harvested and subjected to total RNA extraction (Qiagen RNeasy Plus kit with column-based DNA removal).

RNA-seq with spiked-in standards. ERCC RNA spike-in mix 1 (Invitrogen) was added to samples based on the cell counts (1 μl of 1/10 dilution to 1 million cells). The ERCC RNA Spike-In Control Mixes used here comprise a set of 92 polyadenylated transcripts that mimic natural eukaryotic mRNAs. On the basis of page 12 of the ERCC manual, we calculated the concentrations of the RNA molecules added to total RNA (that is, number of copies of spiked-in molecules per million cell) (Source Data associated with Fig. 3). It is clear that the standards cover a wide range of copy numbers.

Spiked-in RNA-seq analysis. The spiked-in RNA-seq libraries were subsequently sequenced on Illumina HiSeq 2000 and then trimmed using 'cutadapt' with TruSeq Indexed Adapters. The sequences of the ERCC synthetic spiked-in RNAs (http://tools.invitrogen.com/downloads/ERCC92.fa) were then added to both mouse and human genomes (genome.fa). The exon reference (http://tools.invitrogen.com/downloads/ERCC92.gtf) has also been added to the UCSC exon reference.

New bowtie indexes were then built and reads were aligned to the newly built genomes using tophat. The RPKM (reads per kilobase of exon per million) was then computed for each gene and synthetic spiked-in RNA using cufflinks. To renormalize the RNA-seq data using spiked-in control, we followed the same procedure as previously recommended[21]. We used a loess regression to renormalize the RPKM values by using only the spiked-in values to fit the loess. The *affy* package in R provides a function, *loess.normalize*, which will perform loess regression on a matrix of values (defined by using the parameter *mat*) and allows for the user to specify which subset of data to use when fitting the loess (defined by using the parameter *subset*). For this application the parameters mat and subset were set as a matrix of all RPKM values and the row indices of the ERCC spiked-ins, respectively. The default settings for all other parameters were used. The result of this was a matrix of

RPKM values normalized to the control ERCC spiked-ins. Source Data associated with Fig. 3 quantitates the fraction of spiked-in tag counts in each RNA-seq library when tag counts were generated using "htseq-count–mode=intersection-nonempty–stranded=no".

ChIP-seq. For p300, we chemically crosslinked and sonicated cells to generate fractionated genomic DNA. ChIP was performed by using anti-p300 (sc-585, Santa Cruz Biotechnology). The DNA fragments were blunt-end ligated to the Illumina adaptors, amplified, and sequenced by using the Illumina Genome Analyzer II (Illumina). Sequence reads of 25 or 36 bp were obtained by using the Illumina Analysis Pipeline. Publically available ChIP-seq data sets are listed in Source Data associated with Fig. 1 and were obtained from several published studies[12–15,25–28].

ChIP-seq analysis. ChIP libraries were sequenced for 36 or 50 cycles on an Illumina Genome Analyzer II or HiSeq 2000, respectively, according to the manufacturer's instructions. ChIP libraries were aligned to mm9 or hg19 reference genomes using bowtie2 with bowtie indexes derived based on UCSC annotations and Phred+33 selected for qualities. Source Data associated with Fig. 1 summarizes ChIP-seq data sets generated or used for this study along with relevant antibodies. Peak calling for all transcription factors and p300 binding was performed by macs14 (ref. 29) using P value $= 1 \times 10^{-7}$. The control library for all peak-calling libraries was the input DNA performed under T$_H$0 condition. Peaks with false discovery rate (FDR) values more than 30% were further excluded. Peak intensities ('tags' column) were normalized as tags-per-million reads in the original library. Peak calling for H3K27ac libraries was performed using SICER[30] where the window size = 200 bp, gap size = 200 bp and E value = 200. To visualize and normalize ChIP-seq libraries on the UCSC genome browser, we used 'bedtools genomecov' to 'scale' the bam files to tags-per-million values. Furthermore, 'wigToBigWig' was used to generate bigwig files. y-Axis in all gene tracks is in tags per million (TPM).

Delineation of SEs and typical enhancers TEs. To accurately delineate SE domains, we followed the same approach that was proposed earlier[2–4]. We first merged genomic regions within 12.5 kb of one another (using mergeBed in bedtools). We then ranked all regions in a cell type by increasing total ChIP-seq occupancy of p300 or H3K27 acetylation, scaled the data such that the x and y axes were from 0–1 by normalizing to the largest value, and plotted the intensity of ChIP-seq (Fig. 1a). These plots revealed a clear point in the distribution of enhancers where the occupancy signal began increasing rapidly. To geometrically define this point, we found the x-axis point for which a line with a slope of 1 was tangent to the curve. As suggested by Young and colleagues, we defined genomic regions above this point to be SEs. All genomic regions below that point that did not harbour promoters (±5 kb of RefSeq transcription start sites) were then referred to as TEs. The single map of SEs in CD4$^+$ T cells was constructed by merging maps of T$_H$1, T$_H$2 and T$_H$17 SEs (unionBedGraphs). Similarly, TEs in each lineage were delineated as described and then merged in different lineages to build one map for TEs. Since SEs in one lineage can be TEs in other lineages, SE coordinates were then excluded from the final TE map for CD4$^+$ T cells. Source Data associated with Figs 1 and 4 summarize the coordinates of SEs in both human and mouse in our study.

Delineation of cell-type specific SEs. To define cell-type-specific and shared SE domains, we started from the merged map of SEs in T$_H$1, T$_H$2 and T$_H$17 cells (Source Data associated with Fig. 1). We then used 'bedtools intersect' with –f 0.1, 0.3, 0.5 or 0.7 with –a being the coordinate of merged map and –b being the SE coordinates in the corresponding condition and reporting –c in the output (for each entry in A, report the number of overlaps with B and reporting 0 for A entries that have no overlap with B) (Fig. 1b and Extended Data Fig. 1b). We used the pheatmap function to demonstrate the shared and unique SEs based on the outputs of 'bedtools intersect' for the three cell types. Figure 1b corresponds to $f = 0.1$.

Characterizing SE- and TE-associated genes. SE- and TE-associated genes were defined based on the closest genes to these genomic regions (bedtools closest) using RefSeq coordinates of genes. As described in this package, closestBed first searches for features in B (gene coordinates) that overlap a feature in A (SE coordinates). If overlaps are found, gene coordinates that overlap the highest fraction of SE regions are reported. Then in the case of multiple genes overlapping SEs, the gene with the highest fraction of overlap is reported. If no overlaps are found, closestBed looks for the feature in B that is closest (that is, least genomic distance to the start or end of A) to A.

Transcription at T$_H$-specific SE genes. We delineated SE genes as described earlier for T$_H$1, T$_H$2 and T$_H$17 p300 binding (Fig. 1d). We defined a gene to be specific to a lineage if that gene was not present in SE-associated genes in the other two lineages. We then showed the log$_2$ RPKM values for this list of genes across three different lineages. P values were calculated using the Wilcoxon rank-sum test.

Characterization of long ncRNAs associated with SEs. The list of transcribed ncRNAs in T cells was compiled from Hu *et al.*[11]. Hu *et al.*[11] performed the following steps for the identification of ncRNA clusters: (1) call RNA-seq read enriched islands from intergenic regions using SICER (window = 100 bp, gap = 200 bp, E value = 100); (2) keep islands shared by duplicates; (3) pool islands from all

samples, independently done for data sets from total RNA-seq and from PolyA$^+$ RNA-seq; (4) cluster neighbouring islands based on similarity in expression profiles across different samples ($r > 0.8$). Transcribed regions that overlapped SEs were identified using the countOverlaps function in the GenomicRanges package in R (Fig. 1e). To quantitate the correlation levels in transcripts across different T-cell lineages and time points, we used the 'cor' function in R with 'pearson' as 'method' (no $\log_2$ transformation was performed prior to the calculation of correlation) (Fig. 1f, g). Transcript levels for polyA RNAs used for this analysis were extracted from the supplementary table provided in ref. 11. Genomic coordinates of these two groups of ncRNAs are provided in Source Data associated with Fig. 1.

Cumulative distribution of ncRNAs with and without SEs. We used the 'rowSds' function from the library 'matrixStats' in R to calculate the standard deviation in each row for expression levels of ncRNAs with and without SEs. We used ggplot and stat_ecdf() to plot the cumulative distribution of standard deviation in these two groups of ncRNAs. Cumulative distribution in Fig. 1h shows quantitative shift in standard deviation of transcript levels for ncRNAs with SEs relative to those without SEs (P value = 1.326×10^{-7}, Kolmogorov–Smirnov test).

Profile of transcription factor binding at SE genomic regions. To plot the normalized tags-per-million transcription factor binding at SEs and their flanking 40 kb regions, we used the 'ngs.plot.r' package[31] (for example, Fig. 2a). To generate the enrichment of transcription factors at T_H-preferred SEs, we started by counting all tags in .bed files for each transcription factor binding using "bedtools coverage –counts" across the one map of SEs in T cells ($T_H1/T_H2/T_H17$ merged). Furthermore, in Fig. 2d we selected the T_H (1, 2, 17)-preferred SEs as genomic regions identified based on overlapping fraction = 0.1 identified in Fig. 1b. Extended Data Fig. 2a was generated by using ngs.plot on the same set of cell-type-specific coordinates. The normalization has been done as described previously[31]: the coverage data were normalized in two steps. In the first step, the coverage vectors were normalized to have equal length using spline fit. In this case, a cubic spline is fit through all data points and values are taken at equal intervals. This first step of length normalization leads to regions of variable sizes to have equal lengths and is particularly useful for custom regions. The purpose of the second step is to normalize vectors against the corresponding library size—that is, the total read count.

Profile of transcription-factor binding at constituent elements of SEs. We first recovered the original peak regions for p300 binding (constituent enhancers) within SEs from outputs of the peak-calling method (MACS) overlapping SEs/TEs. We then used the HOMER 'annotatePeaks.pl' function to plot the enrichment of transcription factor binding at constituent enhancers in SEs and TEs (Extended Data Fig. 2c).

GO analysis for SE-associated genes. In Fig. 2e, GO enrichment for SE genomic coordinates was carried out using GREAT[24] with default parameters. The top ten terms based on binomial P values were selected in Fig. 2e. In a completely different approach, we characterized the closest genes to SEs. The top GO molecular functions in terms of GSEA "Investigate Gene Sets" were then selected. To calculate the statistical significance of these enrichments, we randomly moved the SE regions around the genome 10^5 times, delineated the closest gene sets to the random genomic domains, and assessed the relative proportion of a gene set that is captured in the actual data versus the shifted SEs. P values for this permutation test are reported in Extended Data Fig. 3a.

GO functional category relevant to cytokine binding is enriched at SE-associated genes in T cells and to a lesser extent in macrophages but not in mouse ES cells and myotubes (Extended Data Fig. 3b). To explore whether 'cytokine binding' is specific to the SE structure in CD4$^+$ T cells, we explored its association within the SE structures of other cell types. The GO molecular function associated with cytokine binding (GO:0019955) was chosen. SE-associated genes in myotubes were used from Whyte *et al.*[4]. SE regions in mouse ES cells and macrophages were chosen based on data sets reported in Source Data associated with Fig. 1. To calculate the statistical significance of this gene category, we shuffled the SE regions of mouse ES cells, macrophages, myotubes and CD4$^+$ T cells around the genome 10^5 times, delineating the gene sets in proximity to the random genomic domains associated with each cell type. We then assessed the relative proportion of the gene set captured in the actual data versus the shifted SEs. P values for this permutation test are reported in the bar graph in Extended Data Fig. 3b.

Analysis of RNA-seq data from *Bach2*-deficient cells. The $\log_2$ fold change of average RPKM values in wild-type and knockout repeats were calculated for SE genes and an equal number of randomly selected TE and other genes in the violin plots (Fig. 3d and Extended Data Fig. 3d). In Extended Data Fig. 3c, d, the RPKM values for the spiked-in measurements were renormalized based on the spiked-in standards. We used ggplot and geom_violin(scale = "area") to plot the impact of loss of BACH2 on gene expression. All genes in SEs, TEs, or the rest of genes were used for the cumulative distribution plots (Fig. 3c and Extended Data Fig. 3c). In Extended Data Fig. 3e, we focused on the top 500 highly expressed genes and explored the effect of BACH2 on three categories among them: genes with SEs (77), with TEs

(125), and without either SEs or TEs (298). Expression levels among these three categories of genes were comparable (Wilcoxon rank-sum test P value = 0.644). However, BACH2 selectively affected highly expressed SE genes in contrast to those with TEs or no enhancers (Kolmogorov–Smirnov test P value = 9.813×10^{-7} and 4.669×10^{-8}).

GSEA plot. The 'gene-set' for GSEA was generated based on genes closest to SEs with minimum 1 RPKM value in any of the three lineages (T_H1, T_H2, T_H17 cells). Three repeats for wild-type and *Bach2*-knockout RNA-seq data were used in the GSEA analysis with default settings (Fig. 3e). The P value for the enrichment was calculated as 0 although -nperm = 10000 was used (with command-line usage of GSEA). In the case of spiked-in GSEA analysis (Extended Data Fig. 4c), two repeats for wild-type and knockouts of renormalized spiked-in data were used.

Pie chart demonstrating *Bach2*-dependent SE genes. BACH2 up- or down-regulated genes (Source Data associated with Fig. 3) were delineated by the 'DEseq' package in R with FDR < 0.05 and fold change > 1.5. Tag counts were calculated using "htseq-count–mode=intersection-nonempty–stranded=no" (Extended Data Fig. 4d). Three repeats of RNA-seq data for wild-type and knockout samples (no spiked-in) were used (Source Data associated with Fig. 1). Direct targets of BACH2 were identified based on BACH2 ChIP-seq data at these two groups of genes. A list of SE genes with at least 1 RPKM expression in T_H1, T_H2 or T_H17 cells was used for this analysis.

Characterization of BACH2-dependent noncoding RNAs. We used "bedtools coverage –counts" to quantitate the enrichment of RNA-seq reads at 501 ncRNAs with SE structure in wild-type and *Bach2*-knockout cells (Source Data associated with Fig. 3). Transcript levels were further normalized to the size of each library (tags per million) and the average of enrichment in three repeats were calculated. Next, we selected ncRNAs with SE structure that were up- or downregulated by BACH2 (>4 fold-change) (Fig. 3f).

Impact of transcription factors on SE- and TE-associated genes. The fold change in RPKM values between wild-type and knockout samples was calculated for SE genes and an equal number of randomly selected TE genes (Extended Data Fig. 4a). For each transcription factor, the difference between SEs and TEs was quantitated using Kullback–Leibler distance between the two distributions for fold changes in the two groups of genes using the KL.dist function in the FNN library in R (Extended Data Fig. 4b). The largest difference between SEs and TEs generated because of loss of BACH2, STAT4 and STAT6 suggests the more selective impact of these transcription factors on SEs.

Pruning SNPs. To ensure that the SNPs associated with disease are in physically independent segments of the genome, we pruned our lists of SNPs (Fig. 4a). Data from the 1000 Genomes (release 20110521) were downloaded from the 1000 Genomes open ftp site. SNPs that were present in each of the six disease conditions were extracted. For each disease, the all-versus-all pairwise r^2 values were calculated. Finally, all variants were greedily pruned until no pair had an r^2 value greater than the threshold (0.5). The number of SNPs pruned for each disease and their genomic coordinates can be found in Source Data associated with Fig. 4.

T-cell SEs in human and enrichment of SNPs. Human SEs in T-cell subsets were characterized based on H3K27ac data in T_H1, T_H2 and T_H17 cells (Source Data associated with Fig. 4). The methodology for the delineation of SEs for human T cells was the same as the one described for the mouse data. We referred to the merged map of the T_H1, T_H2 and T_H17 SEs as the single map of SEs in CD4$^+$ T cells (Source Data associated with Fig. 4). The lists of tag SNPs for all traits except RA were extracted from the GWAS catalogue (December 2013) and only those with P values less than 1×10^{-8} were selected. The list of 101 RA SNPs were chosen from the recent meta-analysis of RA GWASs[7]. The percentages of SNPs within SEs/TEs were calculated based on the number of SNPs falling into the genomic domains labelled as SEs/TEs. To account for the size of the genome that these two types of enhancers span, we divided the number of SNPs enriched in SEs/TEs by the total size of SEs (66.5338 Mb) and TEs (63.12915 Mb) and reported the number of SNPs in every 10 Mb of the genome in Fig. 4a. The permutation test for the enrichment P value was calculated by generating 10^6 permutations of SEs and TEs in the genome (excluding unmappable regions in each permutation) and considering the number of iterations where the number of overlapping SNPs with random SEs/TEs exceeded the observed ones in CD4$^+$ T cells. SNPs in linkage disequilibrium with the list of tag SNPs were determined from the 1000 Genomes Project using $r^2 = 0.9$ and distance limit = 500 using SNAP toolbox (Extended Data Fig. 5a).

RA risk genes and SEs. The list of 98 rheumatoid arthritis (RA) risk genes was extracted from the study of Plenge and colleagues[7] (Fig. 4b). H3K27ac data for muscle, CD14$^+$ and CD56$^+$ cells are summarized in Source Data associated with Fig. 4.

Quantitating the effect of tofacitinib on different groups of genes. For each donor (except donor 4), the RPKM values with spiked-in were renormalized and the fold changes at SE/TE genes were reported (Fig. 4c). No spiked-in was used for

the RNA-seq analysis of donor 4. The P values were calculated based on Wilcoxon signed-rank test (wilcox.test function in R) for violin and box plots. The violin plots used 'scale = 'area''. In Extended Data Fig. 4c, for each donor, the top 100 highly expressed genes in non-treated RNA-seq data were selected and categorized as having SEs or not.

IBD, MS and T2D risk genes and SEs. The candidate genes associated with RA[7], inflammatory bowel disease (IBD)[22], multiple sclerosis (MS)[23] and type 2 diabetes (T2D)[32] were chosen based on a recent meta-analysis of GWAS data. More than half of RA risk genes (53/98) accommodated SEs in CD4$^+$ T cells. In line with the enrichment of SNPs associated with IBD and MS in T-cell SEs (Fig. 4a), around half of IBD (91/216) and MS risk genes (36/87) were associated with SEs in T cells. In contrast, T2D risk genes showed little association with SEs (4/65) (Fisher's exact test, P value = 0.4). RA and IBD risk genes with SEs are selectively targeted by a JAK inhibitor, tofacitinib. Cumulative plots depict the fold change in expression (log$_2$) after 0.3 μM tofacitinib treatment of human CD4$^+$ T cells at RA (Extended Data Fig. 6b), IBD (Extended Data Fig. 6c) and MS (Extended Data Fig. 6d) risk genes with or without SEs.

25. Ghisletti, S. *et al.* Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32,** 317–328 (2010).
26. Nakayamada, S. *et al.* Early Th1 cell differentiation is marked by a Tfh cell-like transition. *Immunity* **35,** 919–931 (2011).
27. Wei, G. *et al.* Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity* **35,** 299–311 (2011).
28. Hawkins, R. D. *et al.* Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity* **38,** 1271–1284 (2013).
29. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
30. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25,** 1952–1958 (2009).
31. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15,** 284 (2014).
32. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genet.* **44,** 981–990 (2012).

LETTER

ATG14 promotes membrane tethering and fusion of autophagosomes to endolysosomes

Jiajie Diao[1,2,3,4,5]*, Rong Liu[6,7,8]*, Yueguang Rong[6,7], Minglei Zhao[1,2,3,4,5], Jing Zhang[6,7], Ying Lai[1,2,3,4,5], Qiangjun Zhou[1,2,3,4,5], Livia M. Wilz[9], Jianxu Li[9], Sandro Vivona[1,2,3,4,5], Richard A. Pfuetzner[1,2,3,4,5], Axel T. Brunger[1,2,3,4,5] & Qing Zhong[6,7]

Autophagy, an important catabolic pathway implicated in a broad spectrum of human diseases, begins by forming double membrane autophagosomes that engulf cytosolic cargo and ends by fusing autophagosomes with lysosomes for degradation[1,2]. Membrane fusion activity is required for early biogenesis of autophagosomes and late degradation in lysosomes[3–7]. However, the key regulatory mechanisms of autophagic membrane tethering and fusion remain largely unknown. Here we report that ATG14 (also known as beclin-1-associated autophagy-related key regulator (Barkor) or ATG14L), an essential autophagy-specific regulator of the class III phosphatidylinositol 3-kinase complex[8–11], promotes membrane tethering of protein-free liposomes, and enhances hemifusion and full fusion of proteoliposomes reconstituted with the target (t)-SNAREs (soluble N-ethylmaleimide-sensitive factor attachment protein receptors) syntaxin 17 (STX17) and SNAP29, and the vesicle (v)-SNARE VAMP8 (vesicle-associated membrane protein 8). ATG14 binds to the SNARE core domain of STX17 through its coiled-coil domain, and stabilizes the STX17–SNAP29 binary t-SNARE complex on autophagosomes. The STX17 binding, membrane tethering and fusion-enhancing activities of ATG14 require its homo-oligomerization by cysteine repeats. In ATG14 homo-oligomerization-defective cells, autophagosomes still efficiently form but their fusion with endolysosomes is blocked. Recombinant ATG14 homo-oligomerization mutants also completely lose their ability to promote membrane tethering and to enhance SNARE-mediated fusion in vitro. Taken together, our data suggest an autophagy-specific membrane fusion mechanism in which oligomeric ATG14 directly binds to STX17–SNAP29 binary t-SNARE complex on autophagosomes and primes it for VAMP8 interaction to promote autophagosome–endolysosome fusion.

Endoplasmic reticulum (ER)-derived STX17, SNAP29 and endolysosome-localized VAMP8 converge on the complete autophagosome and mediate autophagosome–lysosome fusion[6,7]. STX17 also recruits ATG14 to the ER–mitochondria contact site and promotes autophagosome formation and/or maturation[12]. However, a potential function of ATG14 in membrane tethering and fusion has not yet been explored.

We first tested whether ATG14 physically interacts with autophagic SNAREs. In an in vitro glutathione S-transferase (GST) pull-down assay with recombinant full-length STX17, SNAP29 and VAMP8 proteins purified from Escherichia coli (Extended Data Fig. 1a), recombinant ATG14 bound to STX17 and SNAP29 but not to VAMP8 (Fig. 1a). STX17 G244/248L, an autophagosome targeting-deficient mutant[6], still bound to ATG14 (Fig. 1a). Recombinant ATG14 bound to STX17 alone and the STX17–SNAP29 binary t-SNARE complex, but not to the STX17–SNAP29–VAMP8 ternary complex (Fig. 1b), suggesting that ATG14 binds before formation of trans-SNARE complex or that it can interact with partly folded trans-SNARE complex (but not fully folded cis-SNARE

complex). Overexpression of ATG14 stabilized the STX17–SNAP29 binary complex in a co-immunoprecipitation assay (Extended Data Fig. 1b). Deletion mapping analysis showed that the coiled-coil domain (CCD) of ATG14 interacts with the SNARE core domain of STX17 (Extended Data Fig. 1c, d).

We also tested whether STX17 is a part of the beclin 1/class III phosphatidylinositol 3-kinase (PI3KC3) complex. ATG14 co-fractionated with STX17 and beclin 1 in a 669-kilodalton (kDa) complex separated by size-exclusion chromatography (SEC) (Extended Data Fig. 1e), and ATG14 interacted both with beclin 1 and with STX17. However, STX17 only
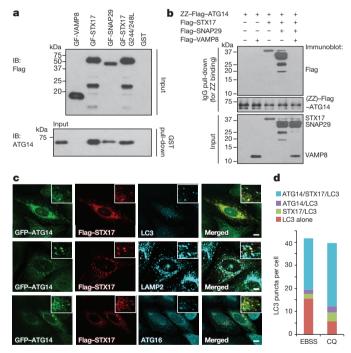

Figure 1 | ATG14 interacts with STX17–SNAP29 on mature autophagosomes. a, Interaction between purified recombinant ATG14 and GST–Flag-tagged (GF) autophagic SNAREs using an in vitro GST pull-down assay followed by western blot (IB, immunoblot). **b**, Recombinant ATG14 binds to STX17–SNAP29 binary t-SNARE complex but not STX17–SNAP29–VAMP8 ternary complex using an in vitro immunoglobulin-G (IgG) pull-down assay followed by western blot. **c**, Flag–STX17, GFP–ATG14, endogenous LC3, LAMP2 and ATG16 were detected by immunostaining in EBSS-starved U₂OS cells ($n = 20$). Scale bars, 5 μm. **d**, Corresponding statistical analysis of co-localization of ATG14, STX17 and LC3 upon treatment with EBSS or chloroquine (CQ).

[1]Department of Molecular and Cellular Physiology, Stanford University, Stanford, California 94305, USA. [2]Department of Structural Biology, Stanford University, Stanford, California 94305, USA. [3]Department of Photon Science, Stanford University, Stanford, California 94305, USA. [4]Department of Neurology and Neurological Sciences, Stanford University, Stanford, California 94305, USA. [5]Howard Hughes Medical Institute, Stanford University, Stanford, California 94305, USA. [6]Center for Autophagy Research, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [7]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [8]College of Food Science & Nutritional Engineering, China Agricultural University, Beijing 100083, China. [9]Division of Biochemistry, Biophysics and Structural Biology, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA.
*These authors contributed equally to this work.

bound to ATG14, but not beclin 1 and vice versa, in immunoprecipitation assays (Extended Data Fig. 1f, g). In cultured U$_2$OS cells, ATG14 co-localized with STX17 and LC3, at least partly, on mature autophagosomes adjacent to or overlapping with LAMP2-labelled autolysosomes upon treatment with Earle's balanced salt solution (EBSS) (Fig. 1c, d). ATG14 also localized to mature autophagosomes with STX17 but not with ATG16 upon treatments with chloroquine or bafilomycin A1 (Extended Data Fig. 2). Together, these results suggest that ATG14 physically and physiologically interacts with the STX17–SNAP29 binary t-SNARE complex on mature autophagosomes.

We then tested whether ATG14 participates directly in membrane tethering in a single-vesicle/liposome assay[13,14] (Methods and Extended Data Fig. 3a). Recombinant ATG14 alone was sufficient to strongly promote protein-free liposome tethering (Fig. 2a). However, lipid-mixing of protein-free liposome membranes was not promoted by ATG14 (Extended Data Fig. 3b and Fig. 2b). The membrane-tethering activity of ATG14 required its membrane-binding Barkor/ATG14(L) autophagosome targeting sequence (BATS) domain[15], but the BATS domain alone was insufficient for membrane tethering (Extended Data Fig. 3c, d). The membrane-tethering activity of ATG14 is not solely due to membrane curvature sensing, since no membrane-tethering activity was observed using another curvature sensing ALPS motif-containing ARFGAP1 protein[16] or the membrane curvature-inducing protein BIF-1/endophilin B1 (ref. 17) (Extended Data Fig. 3e, f). ATG14 membrane-tethering activity was independent of phosphatidylinositol 3-phosphate (PI3P) for small (50 nm) liposomes, but increased in the presence of PI3P for large (400 nm) liposomes (Extended Data Fig. 3g, h). Taken together, ATG14 alone is a membrane tether.

It is unclear whether autophagic SNAREs form a fusion-competent α-helical bundle and possess fusogenic activity. We co-expressed and co-purified the four SNARE core domains of VAMP8 (10–74), STX17 (164–227) and SNAP29 (39–116, 194–258), and determined the crystal structure of the complex at 1.4 Å resolution (Methods, Extended Data Fig. 4a–c and Extended Data Table 1). The autophagic SNARE complex forms a parallel four α-helix bundle (Fig. 2c). At the centre of the complex is a conserved ionic layer consisting of R37 from VAMP8, Q196 from STX17 and Q84/Q230 from SNAP29 that is characteristic for all

SNARE complexes. The overall structure of the autophagic SNARE complex is similar to that of the neuronal, yeast and endosomal SNARE complexes determined so far (Extended Data Fig. 4c).

We tested the fusogenic activity of autophagic SNAREs using protein-reconstituted proteoliposome ensemble lipid- and content-mixing assays (Methods). Proteoliposomes reconstituted with autophagic SNAREs (mean diameter 50 nm) underwent lipid-mixing at a relatively low level, which was greatly enhanced by the addition of ATG14 (Fig. 2d). ATG14 also enhanced content-mixing in the presence of autophagic SNAREs (Fig. 2e), using an ensemble content-mixing assay with the content marker sulphorhodamine B that was used in previous studies[18] (Methods and Extended Data Fig. 5a). Examination of membrane morphologies by cryo-electron microscopy revealed that hemifusion diaphragms were increased when ATG14 was added to proteoliposomes reconstituted with autophagic SNAREs (Fig. 2f, g). Thus, ATG14 enhances hemifusion and complete fusion of proteoliposomes reconstituted with autophagic SNAREs.

The ATG14 fusion-enhancing effect is specific to autophagic SNAREs, since ATG14 did not enhance membrane fusion mediated by neuronal SNAREs (Extended Data Fig. 5b, c). We tested whether the ATG14 fusion-enhancement is due to its tethering activity or its interaction with autophagic SNAREs. We observed that an STX17 binding-deficient mutant (CCD deletion) of ATG14 retained its membrane-tethering function but lost its fusion-enhancing activity (Extended Data Fig. 5d, e), suggesting that the interaction between ATG14 and STX17 (as part of the autophagic t-SNARE or trans-SNARE complex) contributes to the fusion enhancement.

SEC coupled with multi-angle static light scattering (SEC–MALS) experiments indicated that recombinant ATG14 exists in monomeric, dimeric and tetrameric states (Extended Data Fig. 5f); upon separating different oligomeric forms, we found that the dimeric form of ATG14 is more active than the monomeric form in membrane tethering (Extended Data Fig. 5g). We confirmed ATG14 self-interaction in a co-immunoprecipitation assay (Fig. 3a). In the presence of the crosslinker disuccinimidyl suberate (DSS), we found that the amino terminus of ATG14 but not the CCD domain was required for homo-oligomerization (Extended Data Fig. 5h, i). Oligomeric forms of ATG14 could also be observed on
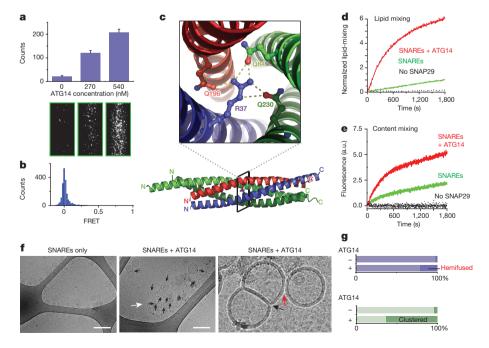

Figure 2 | ATG14 promotes membrane tethering and enhances autophagic SNARE-mediated fusion. a, Purified recombinant ATG14 promotes protein-free single liposome tethering (Methods and Extended Data Fig. 3a). Top: mean number of tethered liposomes ($\pm$ s.d.) ($n = 15$) at random locations in the sample chamber; bottom: corresponding representative images ($n = 15$). **b,** Fluorescence resonance energy transfer (FRET) efficiency profile between single donor/acceptor-dye liposome pairs upon addition of recombinant ATG14 (Methods and Extended Data Fig. 3b). **c,** The crystal structure of the autophagic SNARE complex is shown at the bottom and a close-up view of the ionic layer at the centre is shown at the top. **d,** ATG14 enhances ensemble lipid-mixing of proteoliposomes reconstituted with autophagic SNAREs ($n = 3$). **e,** ATG14 enhances ensemble content-mixing of proteoliposomes reconstituted with autophagic SNAREs ($n = 3$); a.u., arbitrary units. **f,** Representative cryo-electron micrographs ($n = 20$) of proteoliposomes reconstituted with autophagic SNAREs. Left: without ATG14. Middle: with ATG14; black and white arrows indicate hemifusion diaphragms and proteoliposome clusters, respectively. Right: close-up view; black and red arrows indicate a hemifused diaphragm and a tethered proteoliposome pair, respectively. Scale bars, 200 nm. **g,** Percentage of hemifused or clustered proteoliposomes micrographs ($n = 20$).
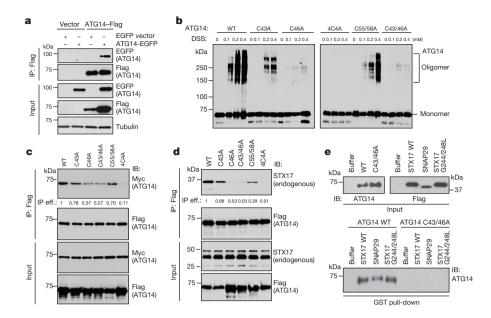
Figure 3 | ATG14 homo-oligomerization is required for autophagic SNARE binding. **a**, Immunoprecipitation (IP) of ATG14–eGFP by ATG14–Flag transfected into HEK293T cells. **b**, Mapping ATG14 oligomerization sites to its cysteine repeats by cross-linking with DSS and analysed by SDS–polyacrylamide gel electrophoresis (SDS–PAGE). **c**, Interaction between Flag-tagged wild-type (WT) ATG14 and mutants with Myc-ATG14 observed in a co-immunoprecipitation assay. IP eff., immunoprecipitation efficiency. **d**, Interaction between Flag-tagged WT ATG14 and mutants with endogenous STX17 observed in a co-immunoprecipitation assay. The immunoprecipitation efficiency (immunoprecipitation/input) in **c** and **d** was normalized by the ratio of immunoprecipitated Flag–ATG14 or interactants versus their inputs. **e**, *In vitro* GST pull-down of purified recombinant WT ATG14 or C43A/C46A mutant with GST-tagged autophagic SNAREs followed by western blot.

native gels without crosslinking and were sensitive to reducing agents (Extended Data Fig. 5j). Upon rapamycin treatment, ATG14 oligomerization was increased compared with unstressed conditions (Extended Data Fig. 5k).

Four evolutionarily conserved cysteine repeats are localized to the N terminus of ATG14 and are essential for its ER localization and autophagy function[19]. We found that C46 is mainly responsible for ATG14 homo-oligomerization. C43 had a minor effect, while C55 and C58 had little impact on homo-oligomerization of ATG14 (Fig. 3b). The mutants C43A/C46A and C43A/C46A/C55A/C58A (abbreviated as 4C4A) showed less self-interaction in a co-immunoprecipitation assay (Fig. 3c); we

refer to these mutants collectively as homo-oligomerization-deficient (HOD) mutants. Remarkably, ATG14 HOD mutants almost completely lost their ability to interact with the autophagic SNARE protein STX17 in a co-immunoprecipitation assay (Fig. 3d) and an *in vitro* pull-down assay (Fig. 3e). Thus, ATG14 homo-oligomerization is essential for its interaction with autophagic SNAREs.

The interaction between these ATG14 HOD mutants and beclin 1 remained intact (Extended Data Fig. 6a). In a reconstituted system purified *in vitro*, ATG14 enhanced the lipid kinase activity of the core VPS34–p150–beclin 1 complex (Extended Data Fig. 6b–e). Recombinant ATG14 HOD mutants could still enhance the lipid kinase activity catalysed by
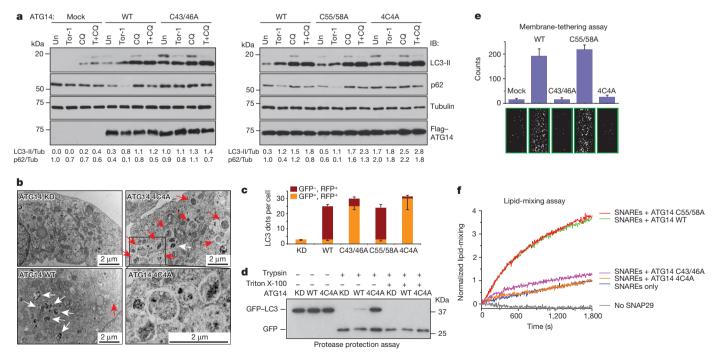

Figure 4 | ATG14 homo-oligomerization is required for autophagosomal fusion with endolysosomes *in vivo* and *in vitro*. a, Autophagic flux analysis in ATG14-deficient U2OS cells reconstituted with WT ATG14 and mutants. Un, untreated; Tor-1, Torin 1. **b**, Accumulation of complete double-membrane autophagosome (red arrows) under TEM in ATG14 4C4A mutant but not in ATG14 WT or knockdown (KD) cells treated with rapamycin (n = 15). White arrows denote autolysosomes. **c**, Quantitative analysis of acidified autophagosomes (GFP−RFP+) versus neutral autophagosomes (GFP+RFP+)

per cell in rapamycin-treated ATG14 knockdown U2OS cells expressing WT or mutant ATG14 transfected with mRFP–GFP–LC3 (mean ± s.d.) (n = 20). **d**, GFP–LC3 was protected from trypsin digestion in ATG14 4C4A mutant but not in ATG14 WT or knockdown cells. **e**, Tethering of protein-free liposomes by recombinant WT ATG14 and specified mutants using the single-vesicle membrane-tethering assay (mean number of tethered liposomes ± s.d.) (n = 20). **f**, Ensemble lipid-mixing activity of WT ATG14 and mutants between proteoliposomes reconstituted with autophagic SNAREs (n = 3).

the VPS34–p150–beclin 1 core complex, but not as efficiently as wild-type (WT) ATG14 (Extended Data Fig. 6f, g). In addition, ATG14 HOD mutants could still localize to autophagosomes (Extended Data Fig. 7) and ER–mitochondria contact sites (Extended Data Fig. 8a), suggesting these membrane localization events are largely dispensable of ATG14 homo-oligomerization.

Furthermore, we investigated whether ATG14 oligomerization is important for autophagosomal fusion with endolysosome *in vivo*. We complemented ATG14-depleted U$_2$OS cells[8] with either WT ATG14 or ATG14 HOD mutants. The autophagy flux was blocked in cells expressing the ATG14 HOD mutants but not in cells expressing either WT ATG14 or the C55A/C58A mutant (Fig. 4a). Furthermore, in cells expressing ATG14 4C4A mutant, mature autophagosomes (LC3$^+$ and STX17$^+$) but not phagophores (Atg16$^+$) accumulated (Extended Data Fig. 8b–d). Accumulation of sealed double-membrane autophagosomes in ATG14 4C4A mutant cells was detected by electron microscopy (Fig. 4b). In addition, we measured autophagosome maturation using LC3 tagged with acid-resistant monomeric red fluorescent protein (mRFP) and acid-sensitive green fluorescent protein (GFP)[20]. Autophagosome acidification/maturation was dramatically reduced in cells expressing ATG14 HOD mutants compared with WT ATG14 (Fig. 4c and Extended Data Fig. 9a). In a modified protease protection assay that was previously used in yeast studies[21] (Extended Data Fig. 9b, c), more GFP–LC3 was protected from protease digestion in cells expressing ATG14 4C4A mutant than that in cells expressing WT ATG14 (Fig. 4d). Finally, we tested whether ATG14 homo-oligomerization is essential for its roles in membrane tethering and enhancement of SNARE-mediated fusion *in vitro*. Purified recombinant ATG14 HOD mutants lacked membrane-tethering activity (Fig. 4e) and lost their ability to enhance ensemble lipid mixing of proteoliposomes reconstituted with autophagic SNAREs (Fig. 4f). Together, all these results demonstrate a crucial role of ATG14 homo-oligomerization in regulating membrane fusion between autophagosome and endolysosomes.

In this study, we observed that, in addition to its localization to phagophore[8–11], ATG14 also localizes to mature autophagosomes. It is still not clear how ATG14 is recruited to mature autophagosomes. Most of ATG14 (or the autophagy-specific class III PI3K complex) dissociates from the phagophore membrane together with other ATGs during autophagosome formation[22]. It is possible that a small fraction remains on the membrane. Alternatively, after dissociation from the phagophore membrane, ATG14 (probably without other PI3KC3 components) is re-recruited to mature autophagosomes to facilitate the autophagosome–lysosome fusion. On autophagosomes, ATG14 promotes and stabilizes the STX17–SNAP29 binary t-SNARE complex assembly, and primes it for interaction with endolysosome-localized VAMP8 to promote membrane fusion between autophagosomes and endolysosomes. The membrane-tethering activity of ATG14 probably stabilizes fusion intermediates and facilitates fusion progression. This fusion process is probably tightly regulated and requires other protein factors[23–25]. Dissecting this regulated machinery will be essential to understand the molecular mechanisms of autophagy, and it will facilitate the development of therapeutics specifically to modulate autophagy in human diseases resulting from its dysregulation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Levine, B. & Klionsky, D. J. Development by self-digestion: molecular mechanisms and biological functions of autophagy. *Dev. Cell* **6,** 463–477 (2004).
2. Huang, W.-P. & Klionsky, D. J. Autophagy in yeast: a review of the molecular machinery. *Cell Struct. Funct.* **27,** 409–420 (2002).
3. Moreau, K., Ravikumar, B., Renna, M., Puri, C. & Rubinsztein, D. C. Autophagosome precursor maturation requires homotypic fusion. *Cell* **146,** 303–317 (2011).
4. Nair, U. *et al.* SNARE proteins are required for macroautophagy. *Cell* **146,** 290–302 (2011).
5. Ragusa, M. J., Stanley, R. E. & Hurley, J. H. Architecture of the Atg17 complex as a scaffold for autophagosome biogenesis. *Cell* **151,** 1501–1512 (2012).
6. Itakura, E., Kishi-Itakura, C. & Mizushima, N. The hairpin-type tail-anchored SNARE syntaxin 17 targets to autophagosomes for fusion with endosomes/lysosomes. *Cell* **151,** 1256–1269 (2012).
7. Takats, S. *et al.* Autophagosomal Syntaxin17-dependent lysosomal degradation maintains neuronal function in Drosophila. *J. Cell Biol.* **201,** 531–539 (2013).
8. Sun, Q. *et al.* Identification of Barkor as a mammalian autophagy-specific factor for Beclin 1 and class III phosphatidylinositol 3-kinase. *Proc. Natl Acad. Sci. USA* **105,** 19211–19216 (2008).
9. Itakura, E., Kishi, C., Inoue, K. & Mizushima, N. Beclin 1 forms two distinct phosphatidylinositol 3-kinase complexes with mammalian Atg14 and UVRAG. *Mol. Biol. Cell* **19,** 5360–5372 (2008).
10. Matsunaga, K. *et al.* Two Beclin 1-binding proteins, Atg14L and Rubicon, reciprocally regulate autophagy at different stages. *Nature Cell Biol.* **11,** 385–396 (2009).
11. Zhong, Y. *et al.* Distinct regulation of autophagic activity by Atg14L and Rubicon associated with Beclin 1-phosphatidylinositol-3-kinase complex. *Nature Cell Biol.* **11,** 468–476 (2009).
12. Hamasaki, M. *et al.* Autophagosomes form at ER-mitochondria contact sites. *Nature* **495,** 389–393 (2013).
13. Diao, J. *et al.* A single vesicle–vesicle fusion assay for in vitro studies of SNAREs and accessory proteins. *Nature Protocols* **7,** 921–934 (2012).
14. Diao, J. *et al.* Native alpha-synuclein induces clustering of synaptic-vesicle mimics via binding to phospholipids and synaptobrevin-2/VAMP2. *eLife* **2,** e00592 (2013).
15. Fan, W., Nassiri, A. & Zhong, Q. Autophagosome targeting and membrane curvature sensing by Barkor/Atg14(L). *Proc. Natl Acad. Sci. USA* **108,** 7769–7774 (2011).
16. Bigay, J., Casella, J. F., Drin, G., Mesmin, B. & Antonny, B. ArfGAP1 responds to membrane curvature through the folding of a lipid packing sensor motif. *EMBO J.* **24,** 2244–2253 (2005).
17. Farsad, K. *et al.* Generation of high curvature membranes mediated by direct endophilin bilayer interactions. *J. Cell Biol.* **155,** 193–200 (2001).
18. Kyoung, M. *et al.* In vitro system capable of differentiating fast Ca2+-triggered content mixing from lipid exchange for mechanistic studies of neurotransmitter release. *Proc. Natl Acad. Sci. USA* **108,** E304–E313 (2011).
19. Matsunaga, K. *et al.* Autophagy requires endoplasmic reticulum targeting of the PI3-kinase complex via Atg14L. *J. Cell Biol.* **190,** 511–521 (2010).
20. Kimura, S., Fujita, N., Noda, T. & Yoshimori, T. Monitoring autophagy in mammalian cultured cells through the dynamics of LC3. *Methods Enzymol.* **452,** 1–12 (2009).
21. Nair, U., Thumm, M., Klionsky, D. J. & Krick, R. GFP-Atg8 protease protection as a tool to monitor autophagosome biogenesis. *Autophagy* **7,** 1546–1550 (2011).
22. Koyama-Honda, I., Itakura, E., Fujiwara, T. K. & Mizushima, N. Temporal analysis of recruitment of mammalian ATG proteins to the autophagosome formation site. *Autophagy* **9,** 1491–1499 (2013).
23. Takats, S. *et al.* Interaction of the HOPS complex with Syntaxin 17 mediates autophagosome clearance in Drosophila. *Mol. Biol. Cell* **25,** 1338–1354 (2014).
24. Nakatogawa, H., Ichimura, Y. & Ohsumi, Y. Atg8, a ubiquitin-like protein required for autophagosome formation, mediates membrane tethering and hemifusion. *Cell* **130,** 165–178 (2007).
25. Chen, D. *et al.* A mammalian autophagosome maturation mechanism mediated by TECPR1 and the Atg12-Atg5 conjugate. *Mol. Cell* **45,** 629–641 (2012).

Author Contributions R.L., Y.R., J.Z., L.M.W. and J.L. performed the biological and biochemical experiments characterizing ATG14 function. J.D., Y.L. and R.A.P. performed the *in vitro* membrane tethering and fusion experiments; M.Z. and Q.Z. determined the crystal structure of the autophagic SNARE complex and performed the cryo-electron microscopy experiments; S.V. performed the SEC–MALS experiments, J.D., R.L., A.T.B. and Q.Z. conceived the project, designed the experiments, analysed the data and wrote the manuscript with the help of all authors.

METHODS

Statistical analysis. In general, data shown in column graphs represent the mean ± s.d. or ± s.e.m., as indicated in the figure legends. Sample sizes were chosen after estimating effect size. To determine the group size necessary for adequate statistical power, power analysis was performed using preliminary data sets. Data were analysed for statistical significance after at least three repeated experiments. No data were excluded. The counts were averaged over the 20 images at random locations in each sample. Statistical analysis used Excel (Microscoft).

For the single-vesicle/liposome-tethering assay (Figs 2a and 4e and Extended Data Figs 3c–h and 5d, g), the counts were averaged from at least 15 images at random locations in each sample channel. No spots were excluded. For each set of comparisons between different conditions and/or mutants, the same protein and polyethylene glycol (PEG) surface preparations were used, and the experiments were run in separate channels on the same quartz slide with immobilized liposomes. Although there was some variation in absolute numbers of counts, the relative differences were statistically similar for different experiments. In all panels, error bars are standard deviations from at least 15 random imaging locations in the same sample channel.

Antibodies and cell lines. Antibodies used in this study included anti-ATG14 (#5504, Cell Signaling), anti-Flag M2 (F3165, Sigma), anti-HA (12CA5, Roche), anti-EGFP (GL-8, Clontech), anti-LC3 (7543, Sigma), anti-p62 (PM045, MBL), anti-ATG16 (PM040, M150-3, MBL), anti-STX17 (HPA001204, Sigma), anti-beclin 1 (sc-11427, Santa Cruz), anti-Tom20 (Ab78547, Abcam), anti-LAMP2 (sc-18822, Santa Cruz), anti-Myc (9E10, DSHB) and anti-β-tubulin (E7, DSHB). U2OS and HEK293T were described before[25].

(Proteo-)liposome preparation. The lipids used in this study were phosphatidylcholine (POPC), phosphatidylethanolamine (POPE), biotinylated phosphatidylethanolamine (biotin-PE) and PI3P (Avanti Polar Lipids). Also, 2–4 mol% 1,1′-dioctadecyl-3,3,3′,3′-tetramethylindocarbocyanine perchlorate (DiI) or 1,1′-dioctadecyl-3,3,3′,3′-tetramethylindodicarbocyanine perchlorate (DiD) (Molecular Probes) were used as lipid dyes for the v- and t-proteoliposomes, respectively. The lipid mixtures POPE:POPC:DiI (molar ratio 20:78–76:2–4) or POPE:POPC:PI3P:DiI (molar ratio 20:76–74:2:2–4) were solubilized in chloroform and then dried to form a lipid film on the wall of a glass tube. The same procedures were applied to prepare another aliquot of lipid solution containing the same ratio of PE, PC, DiD, supplemented with 0.1 mol% biotinylated lipid (or with 2% PI3P). The dried lipid film was resuspended in proteoliposome buffer (20 mM HEPES, 100 mM KCl, pH 7.4). After five freeze–thaw cycles, unilamellar proteoliposomes were extruded through polycarbonate filters (50 nm pore size, Avanti Polar Lipids) at least 39 times.

Single vesicle/liposome-tethering assay. All surfaces were coated with PEG to eliminate non-specific binding of liposomes. A PEG-coated quartz slide was assembled into a flow chamber and coated with neutravidin (0.2 mg ml⁻¹). A detailed protocol for this step has been previously described[13]. Briefly, the protein-free acceptor-dye DiD-labelled liposomes were immobilized on the PEG-coated surface during a 30-min incubation at ambient temperature (~25 °C). After two rounds of 200 μl buffer washing, protein-free donor-dye DiI-labelled liposomes (100–200 nM) with or without ATG14 (540 nM, unless specified otherwise) were injected into the flow chamber and placed in a 37 °C incubator for 30–120 min before buffer washing. The incubation time was long enough to reach equilibrium; no detectable increase of spot numbers was observed with a longer incubation time. For the single-vesicle/liposome-tethering assay (Extended Data Fig. 3a), the number of DiI–liposome interactions (tethering) was determined by counting the number of fluorescent spots from DiI dyes in the green detection channel upon excitation at 532 nm (ref. 13). The counts were averaged from at least 15 images at random locations in each sample channel. For each set of comparisons between different conditions and/or mutants, the same ATG14 and PEG surface preparations were used, and the experiments were run in separate channels on the same quartz slide with immobilized DiD-labelled liposomes. This assay was used in Figs 2a and 4e (concentration of ATG14 and mutants: 360 nM) and Extended Data Figs 3c–h and 5d, g.

Single vesicle/liposome FRET-based lipid-mixing assay. The setup and illumination scheme was similar to the single-vesicle/liposome-tethering assay described above, except that FRET to the acceptor (DiD) dyes from the excited donor dyes (DiI) was observed in the red channel (Extended Data Fig. 3b). The donor intensity bleed-through factor (BF) into the acceptor channel was measured as 15% for our setup. The FRET efficiency was then calculated as:

$$\text{FRET efficiency} = \frac{I_{D_{exc}}^{A_{em}} - I_{D_{exc}}^{D_{em}} \times BF}{I_{D_{exc}}^{D_{em}} + I_{D_{exc}}^{A_{em}} - I_{D_{exc}}^{D_{em}} \times BF},$$

where $I_{D_{exc}}^{A_{em}}$ is the ten-frame-averaged intensity value of acceptor dye emission upon excitation of the donor dye, and $I_{D_{exc}}^{D_{em}}$ is the ten-frame-averaged intensity value of donor dye emission upon excitation of the donor dye[13]. This assay was used in Fig. 2b.

SNARE protein reconstitution. SNARE proteins were reconstituted by using the direct method described in ref. 13. Donor-dye and acceptor-dye proteoliposomes were reconstituted with autophagic t-SNAREs (STX17/SNAP29) and v-SNARE (VAMP8), respectively. SNAP29 and STX17 were mixed at a 1.5:1 molar ratio and incubated at 25 °C for 1 h to allow complex formation before reconstitution. The SNARE proteins and proteoliposomes were mixed together at the desired lipid to membrane-anchored protein (L/P) ratio of approximately 100–200. Then the mixture containing ~0.8 wt% octyl glucoside in the buffer was kept at 4 °C for 20 min. The mixture was diluted two times with proteoliposome buffer (20 mM HEPES, 100 mM KCl, pH 7.4), and dialysed against 2 l proteoliposome buffer at 4 °C for overnight. The same protocol was applied for proteoliposomes used in ensemble content-mixing experiments except that no lipid dyes were included and self-quenched sulphorhodamine B (50 mM) was encapsulated in v-SNARE (VAMP8) proteoliposomes[18]. For the experiments with neuronal SNAREs, proteoliposomes were reconstituted with t-SNAREs (syntaxin-1A/SNAP25) at an L/P ratio of 200 and v-SNARE (synaptobrevin-2/VAMP2) at an L/P ratio of 200, both at 0.1 mM lipid concentration.

Ensemble lipid/content-mixing assays. Protein-reconstituted t- and v-SNARE proteoliposomes were mixed at a molar ratio of 1:1. The ensemble lipid-mixing experiments were performed with DiI donor-dye and DiD acceptor-dye labelled t-SNARE and v-SNARE proteoliposomes, respectively, using the protocol described in ref. 26. Briefly, donor dyes were excited with 530 nm laser light. Emission fluorescence intensity was monitored in two channels, at 570 and 670 nm. Lipid mixing was measured as the fluorescence emission (670 nm) of DiD acceptor dyes arising from FRET upon excitation of DiI dyes with 530 nm light.

For the ensemble content-mixing assay, self-quenched sulphorhodamine B molecules encapsulated in v-SNARE proteoliposomes were used as a content indicator[18]. Content mixing was measured by an increase of fluorescence emission at 570 nm of the sulphorhodamine B dyes upon excitation with 530 nm laser light that results as the initially self-quenched dye is diluted upon complete fusion between labelled v-SNARE and unlabelled t-SNARE proteoliposomes.

Fluorescence emission was recorded with a Varian Cary Eclipse model fluorescence spectrophotometer using a quartz cell of 100 μl with a 5 mm path length. All lipid-mixing measurements were performed at 35 ± 2 °C, whereas content-mixing measurements were performed at ambient temperature (~25 °C). The ATG14 concentrations used for the lipid- and content-mixing assays were 1 μM and 360 nM, respectively. The ensemble lipid-mixing assay was used in Figs 2d and 4f and Extended Data Fig. 5c, e. The lipid-mixing traces in these figures were normalized to the value at 1,800 s of the SNAREs-only trace. The ensemble content-mixing assay was used in Fig. 2e.

Cryo-electron microscopy. Proteoliposomes reconstituted with autophagic SNARE proteins at an L/P ratio of 800 were incubated with or without Atg14 (54 nM) at 37 °C for 3 h. Samples were centrifuged at 800g. for 2 min to remove large aggregations. Frozen-hydrated samples were prepared using the procedures described previously[18]. Samples were imaged in low-dose conditions using a TF20 electron microscope (FEI) operating at 200 kV. Images of both conditions were collected at a nominal magnification of ×29,000 and an under-focus of 2.5–3.5 μm on a TemCam-F416 CMOS camera (TVIPS GmbH, 4,096 pixels × 4,096 pixels). Bar graphs of the percentage of hemifused or clustered proteoliposomes (Fig. 2g) were obtained by visual inspection of 20 micrographs, using 1,079 and 661 observed proteoliposomes without and with ATG14, respectively.

Insect cell recombinant protein purification. WT and mutant ATG14 proteins were purified from sf9 cells as described before[15,27]. Briefly, pFASTBACs containing the desired protein sequence were transformed into DH10α-competent cells to purify bacmid DNA. White colonies yielding bacmid DNA were selected from agar plates (50 μg ml⁻¹ kanamycin, 7 μg ml⁻¹ gentamicin, 10 μg ml⁻¹ tetracycline, 100 μg ml⁻¹ Blue-gal, 40 μg ml⁻¹ IPTG). Baculovirus generated from cells containing bacmid DNA were then used to infect sf9 cells to reach an eventual culture volume of 1.0 l. Cells were harvested by centrifugation, resuspended in hypotonic buffer (20 mM Tris-HCl, pH 7.5, 5 mM KCl, 1 mM MgCl2, 1 mM DTT, protease inhibitors) on ice for 15 min, followed by lysis by using a dounce homogenizer with a pestle (Wheaton) 30 times. High-salt buffer (20 mM Tris-HCl, pH 7.5, 250 mM NaCl, 0.5 mM DTT, 20 mM Glycerol) was added at a 1:1 volume ratio to return the lysate to physiological salt levels. Cell lysate was first incubated with IgG beads to allow conjugation of the ZZ tag to beads. Bound protein was eluted using tobacco etch virus (TEV) protease cleavage. Protein purity levels were assessed with SDS–PAGE and Coomassie staining using standard protocols.

Full-length STX17, VAMP8 and SNAP29 expression and purification. All DNAs encoding human proteins were expressed and purified according to the detailed protocol previously reported[28] with modifications. Briefly, WT STX17, VAMP8 and SNAP29 were expressed individually as N-terminal GST-fused proteins from pGEX-4T-1 (GE Healthcare) in BL21 (DE3) at 25 °C. After binding to Glutathione Sepharose 4 Fast Flow beads (GE Healthcare), the proteins were eluted by overnight

cleavage with TEV protease. In a separate purification scheme, WT His-tagged SNAP29 was co-expressed with STX17 from pETDuet-1 (Novagen) in BL21 (DE3) at 25 °C, and His-tagged VAMP8 was purified in the system separately. His6-tagged SNAP29 and co-expressed STX17 were then extracted from membrane fractions in the presence of 110 mM octyl glucoside and purified using Ni^{2+}-NTA agarose (Qiagen). His6-tag removal from purified SNAP29 was performed by TEV protease cleavage. All these three cleaved proteins (STX17, VAMP8 and SNAP29) were subjected to SEC using a Superdex 200 10/300 column (GE Healthcare) in 110 mM octyl glucoside, 20 mM HEPES, pH 7.4, 300 mM NaCl and 1 mM TCEP, respectively. For the *in vitro* binding assay for ATG14 and autophagic SNAREs, the STX17–SNAP29 binary t-SNARE complex or STX17–SNAP29–VAMP8 ternary complex was assembled *in vitro* and separated by SEC. Their binding to ZZ–Flag–ATG14 was then tested in an IgG pull-down experiment followed by a TEV cleavage assay.

Cloning, expression and purification of the autophagic SNARE complex used for crystallization. The SNARE domains of VAMP8 (10–74) and STX17 (164–227) were cloned into the pACYCDuet-1 vector, with the VAMP8 insert between BamHI and SalI restriction sites containing an engineered TEV protease cleavage site at the N terminus, and with the STX17 insert between NdeI and XhoI restriction sites, respectively. The SNARE domains of SNAP29 (39–116, 194–258) were cloned into the pETDuet-1 vector, with the former fragment inserted between NcoI and SalI restriction sites and the latter fragment inserted between NdeI and XhoI restriction sites, respectively. The two plasmids were co-transformed to *E. coli* BL21 (DE3) cells and expressed at 37 °C using auto-inducing LB medium[28]. After lysis and centrifugation, the cell lysate was loaded onto a 5-ml Ni-NTA agarose column (Qiagen), washed with 100 ml wash buffer (50 mM NaPi, pH 8.0, 300 mM NaCl, 20 mM imidazole), and eluted with elution buffer (wash buffer supplemented with 500 mM imidazole). Fractions containing the autophagic SNARE complex were pooled and concentrated using an Amicon Ultra-15 centrifugal filter with 10-kDa molecular mass cut off (Millipore) to reduce the volume. The complex was then purified on a Superdex 75 16/60 column (GE Healthcare) that was equilibrated with SEC buffer 1 (50 mM NaPi, pH 8.0, 100 mM NaCl, 0.5 mM TCEP). Fractions containing the autophagic SNARE complex were pooled, supplemented with TEV protease, and then dialysed at 4 °C overnight against anion exchange buffer A (50 mM HEPES, pH 7.5, 50 mM NaCl and 0.5 mM TCEP). The TEV protease digested complex was then loaded onto a MonoQ 10/10 anion exchange column (GE Healthcare) and eluted with a linear NaCl gradient up to 500 mM to remove TEV protease. Fractions containing autophagic SNARE complex were pooled and purified on a Superdex 75 10/300 column (GE Healthcare) that was equilibrated with SEC buffer 2 (50 mM HEPES, pH7.5, 150 mM NaCl and 0.5 mM TCEP) for a final purification. The purified autophagic SNARE complex was concentrated to 0.83 mg ml^{-1} on the basis of 280 nm ultraviolet absorption and an extinction coefficient of 22460 M^{-1} cm^{-1}.

Crystallization, data collection and structure determination of the autophagic SNARE complex. Crystals were obtained at 22 °C by the hanging-drop vapour diffusion method in a mother liquor containing 0.1 M Bicine, pH 8.5 and 55% 2-methyl-2,4-pentanediol (MPD). The crystals were allowed to grow for 2 weeks before flash freezing in liquid nitrogen. X-ray diffraction data were collected at the Northeastern Collaborative Access Team beamline 24-ID-C at Advanced Photon Source (Extended Data Table 1) using a beam corresponding to a wavelength of 0.9792 Å. The diffraction data were integrated and scaled using the XDS/XSCALE program packages[29]. The structure was determined by molecular replacement using the program PHASER[30] and a poly-alanine search model derived from the crystal structure of the neuronal SNARE complex (Protein Data Bank accession number 1SFC). The model was built using COOT[31] and was refined using PHENIX[32]. In the final model, all residues (100%) fall into the favoured regions of Ramachandran plot. Coordinates and structure factor amplitudes have been deposited in the Protein Data Bank under accession number 4WY4. In Extended Data Fig. 4, neuronal (1SFC)[33], yeast (3B5N)[34], early endosomal (2NPS)[35] and endosomal (1GL2)[36] SNARE structures and electrostatic potential surfaces are shown in two different orientations, and compared with the autophagic SNARE complex structure.

Expression and purification of full-length neuronal SNAREs. Full-length rat syntaxin-1A and VAMP2 were prepared as previously described[37]. The proteins were expressed separately in *E. coli* with an N-terminal, TEV protease-cleavable, hexa-histidine tag fused to syntaxin using plasmid pJEXPRESS414, and to VAMP2 using plasmid pTEV5 (ref. 38). Protein was expressed overnight at 25 °C in auto-inducing media[28] in *E. coli* strain C43 (ref. 39). Cell pellets from 8 l of culture were suspended in 500 ml of 1× phosphate-buffered saline, 5 mM EDTA and 1 mM phenylmethylsulphonyl fluoride (PMSF) supplemented with Complete Protease Inhibitor Cocktail tablets (Roche), and broken by three passes through a M-110-EH microfluidizer (Mircrofluidics) at 15,000 p.s.i. Inclusion bodies were removed by two consecutive 10 min spins at 15,344*g*. in a JA-14 rotor (Beckman Coulter), and the membrane fraction collected by centrifugation at 125,000*g*. for 2 h in a Ti-45 (Beckman Coulter) rotor. Membranes containing syntaxin-1A were further washed

with a buffer containing 10 mM Tris-H_2SO_4, pH 7.5, 10 mM EDTA, 10% glycerol (w/v), centrifuged at 125,000*g*. for 1 h in a Ti-45 rotor. Membrane pellets were resuspended in 20 mM HEPES, pH 7.5, 500 mM NaCl, and 1 mM TCEP, and 10% glycerol (w/v), and centrifuged for an additional 1 h in the same rotor. Membranes were suspended to a concentration of 5 mg of protein per millilitre in 20 mM HEPES, pH 7.5, 500 mM NaCl, 1 mM TCEP, 10 mM imidazole, and 10% glycerol (w/v), 1 mM PMSF and EDTA-free Complete Protease Inhibitor Cocktail. Dodecylmaltoside (Anatrace) was added to 2%, and after incubation at 4 °C for 1 h, the sample was centrifuged for 35 min at 125,000*g* in a Ti-45 (Beckman Coulter) rotor, and the supernatant loaded onto a 1 ml column of Nickel-NTA agarose (Qiagen). The column was washed with 20 mM HEPES, pH 7.5, 300 mM NaCl, 1 mM TCEP, 20 mM imidazole, 110 mM octyl glucoside and 10% glycerol (w/v), and the proteins were eluted in the same buffer containing 450 mM imidazole and 1 M NaCl. EDTA (1 mM) was immediately added to the pooled fraction, and loaded onto a Superdex 200 HR 10/300 GL (GE Healthcare) that was equilibrated with 20 mM HEPES, pH 7.5, 300 mM NaCl, 1 mM TCEP, 110 mM octyl glucoside, and 10% glycerol (w/v). Protein fractions were pooled, and digested with 0.2 mg ml^{-1} TEV protease for 30 min at ambient temperature, after which the TEV protease had precipitated. TEV was removed by centrifugation at 4,500*g* for 10 min in an Eppendorf model 5804 R tabletop centrifuge (Eppendorf North America). The sample in the supernatant was again digested with 100 mg ml^{-1} TEV protease for 30 min at ambient temperature, after which the reaction was complete and the TEV protease had precipitated. TEV was removed by centrifugation at 4,500*g* for 10 min in the same rotor. Cysteine-free SNAP25 was expressed with an N-terminal TEV protease cleavable hexa-histidine tag from plasmid pTEV5[42]. Proteins were expressed overnight at 25 °C in autoinducing media[32] in *E. coli* strain BL21(DE3) at 30 °C. Cells from 4 l of culture were resuspended in 200 ml of 50 mM sodium phosphate pH 8.0, 300 mM NaCl, 20 mM imidazole (bufferA) supplemented with 1 mM PMSF and four EDTA-free protease inhibitor cocktail tablets (Roche). Cells were lysed by three passes through a Emulsiflex C5 homogenizer (Avestin) at 15,000 p.s.i. The lysate was clarified by centrifugation in the Ti45 rotor (Beckman Coulter) for 1.5 h at 125,000*g*. The supernatant was bound to a 5 ml Nickel NTA column (Qiagen) by flowing the lysate on to the column at 1 ml min^{-1} using the Akta prime (GE Healthcare). The column was washed with 150 ml of buffer A supplemented with 50 mM imidazole and eluted with buffer A supplemented with 350 mM imidazole. Protein-containing fractions were combined, DTT was added to 5 mM, EDTA was added to 1 mM and 150 μg of TEV protease was added to remove the hexa-histidine tag. This mixture was dialysed against buffer B (20 mM HEPES pH 7.5, 100 mM NaCl, 4 mM DTT) overnight at 4 °C. The TEV-cleaved SNAP25 was concentrated in a 15 ml Amicon Ultra centrifugal concentrator with a 10,000 molecular mass cutoff membrane (Millipore) to 5 ml and injected on the Superdex 200 (16/60) column (GE Healthcare) equilibrated in buffer B. Protein-containing fractions were combined, the concentration of SNAP25 measured by absorbance at 280 nm and aliquots were frozen in liquid nitrogen.

Cell culture, cell transfection and cell lysate preparation. 293T and U$_2$OS cells were cultured in DMEM (Gibco) supplemented with 10% FBS (HyClone) and 1% Penicillin-Streptomycin Solution (Gibco). Tet-approved FBS (Clontech) was used for tetracycline-inducible short hairpin RNA cells. Cell transfection was performed using Lipofectin 2000 (Invitrogen) or PEI (Polysciences) according to protocols provided by manufacturers. Whole-cell lysates used for immunoprecipitation and immunoblotting were prepared in tandem affinity purification buffer (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.5% Igepal CA-360, 1 mM NaF, 1 mM Na$_3$VO$_4$, 1 mM EDTA and a protease inhibitor mixture) on ice. Unless used for further immunoprecipitation, lysates were mixed with SDS loading buffer (200 mM Tris-HCl, pH 6.8, 400 mM DTT, 16% β-mercaptoethanol, 8% SDS, 2× loading dye base (Amresco) and 40% glycerol), boiled for 10 min and subjected to SDS–PAGE and western blotting according to standard protocols.

Immunoprecipitation assays. Cells were lysed using tandem affinity purification buffer as described above. Whole-cell lysates (input) were collected after pelleting cellular debris using centrifugation. Lysate was then incubated with a 20:1 ratio of lysate to anti-Flag M2 Affinity Gel (Sigma) for 16 h at 4 °C. Beads were washed three times with tandem affinity purification buffer and then eluted with a 1:10 ratio of 200 μg ml^{-1} Flag peptide to lysate and subjected to SDS–PAGE. Immunoblotting was performed following standard procedures.

Immunofluorescence staining. Cells were transfected according to the protocol described above with noted plasmids or treated with doxycycline either to knock down or to overexpress inducible constructs. Twenty-four hours after transfection, cells were trypsinized and transferred to six-well dishes containing coverslips. After another 24 h, cells were fixed using ice-cold methanol for 4 min at 4 °C. Cells were then washed with PBS and blocked with blocking buffer (2.5% BSA + 0.1% Triton X-100 in PBS) at 25 °C for 2 h. Cells were incubated with primary antibodies at 4 °C overnight, washed with PBS buffer and then incubated with appropriate

secondary antibodies for 2 h at 25 °C. Slides were examined using under a laser scanning confocal microscope (Zeiss LSM 510 META ultraviolet–visible).

Autophagy analysis. Autophagy was induced under starvation or rapamycin treatment. For starvation, cells were washed with PBS three times and incubated with EBSS (Gibco) for 1 h at 37 °C. For rapamycin (LC Laboratories) treatment, cells were incubated with 500 nM rapamycin in complete medium for 16 h at 37 °C. To block autophagy flux, 400 nm bafilomycin A1 (LC Laboratories) was added to complete medium and incubated for 4 h at 37 °C. Autophagy activity was assessed using two approaches: LC3 punctae per cell and LC3-II formation. The amounts of LC3-positive punctae per cell were quantitated manually and averaged. Results are representative of three independent experiments. For LC3-II formation detection, cell lysates were prepared as described above and subjected to standard western blotting protocols using 1:10,000 dilution of the antibody against LC3 (Sigma). Western blotting quantitation was conducted using ImageJ band analysis (NIH), relative LC3-II peak values were normalized to β-tubulin loading control peak values.

Protease protection assay in mammalian cells. Two million human U$_2$OS cells were infected with lentivirus expressing GFP–LC3 for 24 h and treated with chloroquine for 2 h before harvest. The cell pellet was collected by centrifugation at 600g for 5 min and resuspended in 1 ml buffer Q (400 mM sucrose, 200 mM Hepes pH 7.5, 10 mM KCl, 1.5 mM MgCl$_2$, 1 mM EDTA and 1 mM DTT). Resuspended cells were forced to go through a needle of a 25-gauge syringe 30 times. The resulting solution was centrifuged twice at 500g for 10 min to remove the unbroken cells and cells debris. The supernatant fraction was aliquoted into 300 μl per tube. Trypsin was added to each tube to the final concentration at 10 μg ml^{-1}, along with 0.4% Triton X-100 if desired. The reaction mix was incubated at 30 °C for 25 min, and SDS loading dye was immediately added to stop the reaction. The sample was then boiled for 10 min.

SEC–MALS analysis for ATG14 homo-oligomerization. SEC–MALS was performed by injecting 25 μg of purified recombinant ATG14 on a WTC-030S5 column (Wyatt Technology) that was previously equilibrated in 10 mM Tris-HCl, 150 mM NaCl, 1 mM DTT, 1 mM EDTA, pH 8.0 at a flow rate of 0.5 ml min^{-1}. The eluted sample was monitored by ultraviolet absorption at 280 nm (Jasco UV-975 ultraviolet–visible system), light scattering at 658 nm (HELEOS system, Wyatt Technology) and differential refractometry (Optilab system, Wyatt Technology). The data analyses used ASTRA 6.0 software (Wyatt Technology). The protein absolute molecular mass was calculated in ASTRA 6.0 (Wyatt Technology) assuming a dn/dc value of 0.185 ml g^{-1} and a theoretical ultraviolet extinction coefficient value of 1.0 ml mg^{-1} cm^{-1}.

26. Su, Z., Ishitsuka, Y., Ha, T. & Shin, Y. K. The SNARE complex from yeast is partially unstructured on the membrane. *Structure* **16,** 1138–1146 (2008).
27. Sun, Q. *et al.* The RUN domain of rubicon is important for hVps34 binding, lipid kinase inhibition, and autophagy suppression. *J. Biol. Chem.* **286,** 185–191 (2011).
28. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41,** 207–234 (2005).
29. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.* **26,** 795–800 (1993).
30. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
31. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
32. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
33. Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. *Nature* **395,** 347–353 (1998).
34. Strop, P., Kaiser, S. E., Vrljic, M. & Brunger, A. T. The structure of the yeast plasma membrane SNARE complex reveals destabilizing water-filled cavities. *J. Biol. Chem.* **283,** 1113–1119 (2008).
35. Zwilling, D. *et al.* Early endosomal SNAREs form a structurally conserved SNARE complex and fuse liposomes with multiple topologies. *EMBO J.* **26,** 9–18 (2007).
36. Antonin, W. *et al.* A SNARE complex mediating fusion of late endosomes defines conserved properties of SNARE structure and function. *EMBO J.* **19,** 6453–6464 (2000).
37. Diao, J. *et al.* Synaptic proteins promote calcium-triggered fast transition from point contact to full fusion. *eLife* **1,** e00109 (2012).
38. Rocco, C. J., Dennison, K. L., Klenchin, V. A., Rayment, I. & Escalante-Semerena, J. C. Construction and use of new cloning vectors for the rapid isolation of recombinant proteins from *Escherichia coli*. *Plasmid* **59,** 231–237 (2008).
39. Miroux, B. & Walker, J. E. Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J. Mol. Biol.* **260,** 289–298 (1996).

LETTER

Structure of the *E. coli* ribosome–EF–Tu complex at <3 Å resolution by C$_s$–corrected cryo–EM

Niels Fischer[1]*, Piotr Neumann[2]*, Andrey L. Konevega[3,4,5], Lars V. Bock[6], Ralf Ficner[2], Marina V. Rodnina[5] & Holger Stark[1,7]

Single particle electron cryomicroscopy (cryo-EM) has recently made significant progress in high-resolution structure determination of macromolecular complexes due to improvements in electron microscopic instrumentation and computational image analysis. However, cryo-EM structures can be highly non-uniform in local resolution[1,2] and all structures available to date have been limited to resolutions above 3 Å[3,4]. Here we present the cryo-EM structure of the 70S ribosome from *Escherichia coli* in complex with elongation factor Tu, aminoacyl-tRNA and the antibiotic kirromycin at 2.65–2.9 Å resolution using spherical aberration (C$_s$)-corrected cryo-EM. Overall, the cryo-EM reconstruction at 2.9 Å resolution is comparable to the best-resolved X-ray structure of the *E. coli* 70S ribosome[5] (2.8 Å), but provides more detailed information (2.65 Å) at the functionally important ribosomal core. The cryo-EM map elucidates for the first time the structure of all 35 rRNA modifications in the bacterial ribosome, explaining their roles in fine-tuning ribosome structure and function and modulating the action of antibiotics. We also obtained atomic models for flexible parts of the ribosome such as ribosomal proteins L9 and L31. The refined cryo-EM-based model presents the currently most complete high-resolution structure of the *E. coli* ribosome, which demonstrates the power of cryo-EM in structure determination of large and dynamic macromolecular complexes.

Determining the structure of large, dynamic biological macromolecules at a uniformly high resolution provides a challenge both for X-ray crystallography and cryo-EM. Here we have used aberration-corrected cryo-EM in combination with extensive computational sorting to solve
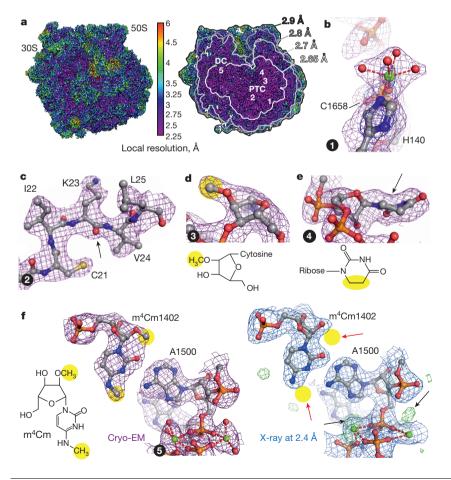

Figure 1 | High-resolution features of the cryo-EM map. a, 3D cryo-EM map of the kirromycin-stalled 70S–EF–Tu–Phe–tRNA^{Phe} complex from *E. coli* coloured according to local resolution. Left, surface view; right, cut-away view. PTC, peptidyl-transferase centre; DC, decoding centre. Numbers (1–5) mark the densities shown in **b–f** rendered at 3σ (purple mesh). **b**, Mg^{2+} ion (green) in octahedral coordination by four water molecules, C1658 of 23S rRNA and His 140 of the protein L3 (dark blue mesh, density at 6σ). **c**, Structure of the protein L14 revealing details such as the thiol group of Cys 21, the branched side chain of Ile 22, the zigzag pattern of Lys 23 side chain, and the carbonyl of the backbone (arrow). **d**, The 2′-O-methyl group (yellow) of Cm2498 in 23S rRNA. **e**, Nucleobase ring of D2449 in 23S rRNA with a characteristic distortion (arrow) of the planar geometry. **f**, The two methyl groups in m^{4}Cm1402 of 16S rRNA (yellow) resolved in the cryo-EM map (left), which are not seen (red arrows) in the X-ray map (right) of the 70S ribosome[9] at 2.4 Å (PDB ID: 4RB5; blue mesh, 2mF$_o$–DF$_c$ map at 1σ). Non-built solvent molecules in the X-ray difference map are indicated by black arrows (green mesh, positive mF$_o$–DF$_c$ map at 3σ).

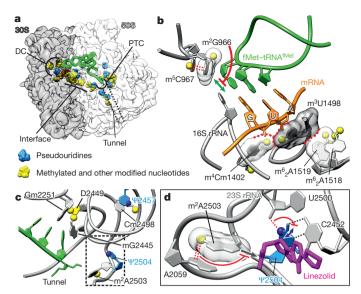[1]3D Electron Cryomicroscopy Group, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. [2]Abteilung Molekulare Strukturbiologie, Institut für Mikrobiologie und Genetik, GZMB, Georg-August Universität Göttingen, Justus-von Liebig Weg 11, 37077 Göttingen, Germany. [3]Molecular and Radiation Biophysics Department, B.P. Konstantinov Petersburg Nuclear Physics Institute of National Research Centre 'Kurchatov Institute', 188300 Gatchina, Russia. [4]St Petersburg Polytechnic University, Polytechnicheskaya, 29, 195251 St Petersburg, Russia. [5]Department of Physical Biochemistry, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. [6]Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. [7]Department of 3D Electron Cryomicroscopy, Institute of Microbiology and Genetics, Georg-August Universität, 37077 Göttingen, Germany.
*These authors contributed equally to this work.

the structure of the *E. coli* 70S ribosome at high resolution to visualize rRNA modifications and dynamic parts of the ribosome. We prepared the 70S *E. coli* ribosome in the codon recognition state with the cognate ternary complex EF-Tu–GDP–Phe-tRNA[Phe] stalled on the ribosome by the antibiotic kirromycin. Currently, a high-resolution crystal structure is available for the 70S–EF-Tu complex from *Thermus thermophilus* at 3.1 Å resolution[6]; however, a comparable structure of the complex from *E. coli* is still lacking, which makes it difficult to integrate structures with the results of biochemical, biophysical and genetic experiments, most of which were obtained with *E. coli* ribosomes. Cryo-EM images were recorded in a 300 kV electron cryo-microscope equipped with a spherical-aberration corrector and a direct electron detector operated in integration mode without alignment of intermediate image frames for motion correction. The aberration corrector was specifically tuned to reduce resolution-limiting aberrations and distortions (Extended Data Fig. 1). We applied a hierarchical classification strategy to sort the ribosome images computationally for the known modes of ribosomal motion and potential sources of heterogeneity in ribosome preparations (Extended Data Fig. 2)[7]. The final cryo-EM map at 2.9 Å overall resolution was subsequently used to refine the atomic model for the entire ribosome, including metal ions (Mg^{2+}, Zn^{2+}) and rRNA modifications (Methods and Extended Data Table 1). The local resolution map[2] is relatively uniform in resolution (Fig. 1a). For large portions of the map the resolution is better than 2.9 Å, whereas only few parts of the ribosome, located at the very periphery, are limited to resolutions >3.5 Å. Locally, average resolutions were 2.8 Å for 65%, 2.7 Å for 44% and 2.65 Å for 24% of the map, as determined by the crystallographic measures FSC$_{work}$ (Fourier shell correlation) and CC$_{work}$ (Pearson correlation coefficient)[8] (Extended Data Fig. 3; see Methods for details). The final 3D map has a similar, or locally even better resolution compared to the available X-ray structures of the *E. coli* ribosome (Fig. 1 and Extended Data Fig. 4a). Also the structural definition of side chains as judged by local real-space correlations between the map and the model is similar to X-ray structures at comparable resolutions (Extended Data Table 2). In the best defined areas of the map Mg^{2+} ions can be visualized along with water molecules in the coordination sphere, indicating an optical resolution of at least 2.8 Å (Fig. 1b).

The present cryo-EM map visualizes for the first time modifications in rRNA, which were not observable in any of the high-resolution X-ray structures of the bacterial ribosome at 2.4–2.8 Å resolution[5,9], and were only seen in high-resolution X-ray structures of the *Haloarcula marismortui* 50S subunit at 2.2–2.4 Å resolution[10] (Extended Data Fig. 4). In contrast, even single methyl groups of nucleosides can be clearly visualized in our cryo-EM map (Fig. 1d, f), as well as the non-planar base of dihydrouridine at position 2449 of 23S rRNA (Fig. 1e), while pseudouridines could be identified indirectly by polar residues within hydrogen-bonding distance of the N1 position. In total, we were able to build all 35 constitutive rRNA modifications[11] of the *E. coli* 70S ribosome.

The rRNA modifications are clustered at the main functional centres of the ribosome (Fig. 2 and Extended Data Fig. 5). Clusters of several rRNA modifications are essential for ribosome function, whereas individual rRNA modifications have an important role in fine-tuning the active centres of the ribosome, as well as in antibiotic resistance and sensitivity[11]. Six individual rRNA modifications at the decoding centre modulate the efficiency and accuracy of translation initiation by stabilizing interactions in the P site[12–14]. The present structure shows that the methyl group of m^{2}G966 in 16S rRNA may act as a 'backstop' for the initiator tRNA, thereby stabilizing base-pairing with the initiation codon (Fig. 2b and Extended Data Fig. 5a; see Extended Data Fig. 5b for definition of rRNA modifications). The binding platform generated by m^{2}G966 is further affected by base stacking interactions with the methyl group of m^{5}C967. Messenger RNA binding to the P site is stabilized by an intricate network of interactions that involves four modified nucleotides. The P-site codon directly contacts m^{4}Cm1402 and m^{3}U1498 in 16S rRNA, which in turn are held in place by the bulky dimethylamine groups on m$^6_2$A1519 and m$^6_2$A1518. The network of long-range

Figure 2 | Modifications of rRNA. a, Clusters of rRNA modifications (yellow and blue) at the functional centres of the ribosome. Initiator fMet-tRNA^{fMet} in the P site is shown in green. PTC, peptidyl-transferase centre; DC, decoding centre. **b**, Close-up of the decoding centre. Methyl group of m^{5}C967 stacks onto the m^{2}G966 base. The methyl group of m^{2}G966 (yellow) restricts the mobility of the initiator tRNA (green). An interaction network of four modified nucleotides stabilizes the mRNA (orange) binding to the P site. **c**, Cluster of rRNA modifications in the peptidyl-transferase centre[17]. The essential Gm2251 forms a base pair with the P-site tRNA (green)[29]; the dashed box indicates the part of the structure shown in **d**. **d**, rRNA modifications impacting antibiotic binding to the peptidyl-transferase centre. The stacking interaction of the A2503 methyl group with the base of A2059 stabilizes A2503 in a conformation compatible with binding of the antibiotic linezolid (magenta, superposition from PDB ID: 3CPW)[30]. The nitrogen in Ψ2504 forms an additional hydrogen bond that stabilizes the nucleotide conformation.

interactions provides the basis for the action of the antibiotic kasugamycin, which binds in the P site and requires dimethylation of m$^6_2$A1519 for its function[15]. In the A site of the decoding centre, the aminoglycoside class of antibiotics directly binds to a monomethylated residue, m^{5}C1407 in 16S rRNA, which is needed for optimum drug activity[16] (Extended Data Fig. 5c).

Assembly and peptidyl-transferase activity of the 50S subunit require a network of six modified nucleosides at the entry site of the peptide exit tunnel (Fig. 2c)[17]. Some of these modifications also modulate antibiotic sensitivity of the ribosome, which can now be explained at a structural level. For instance, resistance against the antibiotic linezolid caused by the loss of methylation of m^{2}A2503[18] can arise from the lack of stabilization by the m^{2}A2503 methyl group on the stacking interaction with A2059, which keeps A2503 in a position that allows the antibiotic to bind. Conversely, the loss of the nitrogen at position 1 in Ψ2504 induces hypersensitivity against the antibiotics linezolid, tiamulin and clindamycin[19], which can be due to the loss of interactions that stabilize Ψ2504 in a conformation less favourable for antibiotic binding (Fig. 2d). The present structure also rationalizes data on 23S rRNA modifications impacting intersubunit bridges and the peptide exit tunnel. For instance, a cluster of three pseudouridines (1911, 1915 and 1917) in helix 69, the central part of intersubunit bridge B2a, forms an enhanced stacking network, explaining their role in stabilizing the helix 69 structure (Extended Data Fig. 5c). The methyl group of m^{2}G1835 has a pivotal role in shaping the junction of four 23S rRNA helices (67–70) which, in turn, directly face 16S rRNA and form intersubunit bridges B2b and B2c (Extended Data Fig. 5d), in line with data showing that this methyl group affects subunit association and cell fitness under stress conditions[20]. A cluster of four modifications (m^{1}G745, Ψ746, m^{5}U747 and m^{6}A1618) lines the peptide exit tunnel (Extended Data Fig. 5e), of which m^{6}A1618 has been shown to be important for cell growth and fitness[21]. Notably,

m[6]A1618 and its direct environment have also been recently predicted as the most promising novel drug target by a computational approach[22].

Generally, obtaining uniform resolution for large macromolecules and especially for bound factors is difficult for cryo-EM[1,2] and X-ray crystallography[23–25], because of mobility and/or occupancy problems. Owing to extensive computational sorting of images, our cryo-EM map is comparable in local resolution variations to crystallographic maps. Higher B factors in X-ray structures are also in line with structural elements exhibiting the highest mobility in molecular dynamics simulations (Extended Data Fig. 6), suggesting that structural dynamics is the prevailing factor limiting local resolution. In X-ray crystallography, those flexible regions may be resolved when stabilized by crystal packing interactions. In cryo-EM, however, computational sorting of images can be employed to improve the local resolution of dynamic structural features. In the cryo-EM map, the local resolution of the EF-Tu–GDP–kirromycin–Phe–tRNA[Phe] complex was improved by sorting for ligand occupancy, resulting in a well-defined density at only slightly lower local resolution compared to the ribosome core. Notably, we find a clear density for the antibiotic kirromycin (Fig. 3a), underlining the power of

computational sorting in cryo-EM. Whereas the overall architecture of the EF-Tu–GDP–kirromycin–Phe–tRNA[Phe] complex is similar to that reported by X-ray crystallography[6,26], there are local differences in the conformation of the distorted tRNA in the A/T state, in the orientation of the catalytic His84 residue and in the interactions between domain 2 of EF-Tu and the 30S subunit (Extended Data Fig. 7). Importantly, a shift in domain 2 (residues 219 to 226) of EF-Tu upon ribosome binding appeared crucial for distortion of the 3′ end of tRNA and, ultimately, GTPase activation[26]. The cryo-EM structure shows a similar distortion of the tRNA, but different interactions of EF-Tu domain 2, indicating that the reported changes in domain 2 may not be essential for the mechanism of catalytic activation. The highly mobile protein L9 is stabilized in crystals in an extended conformation by contacts to a neighbouring ribosome in the crystal lattice (Fig. 3b). In cryo-EM structures, this stabilization is absent and usually only the amino-terminal domain of the protein is structurally well-defined. Nevertheless, we were able to build the complete model for the conformation of L9 on the ribosome in solution, which reveals the contacts of L9 to the 30S subunit (Fig. 3b, c and Extended Data Fig. 8a). Notably, both conformations
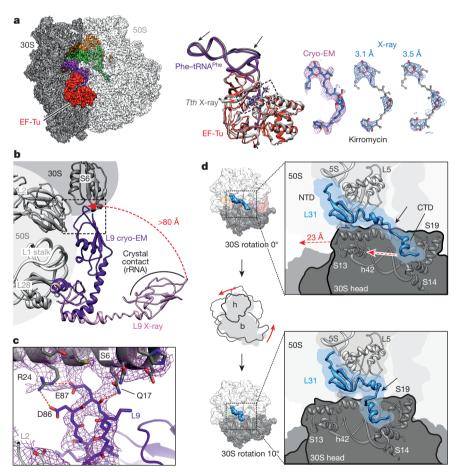

Figure 3 | Structure of ribosome ligands and mobile elements. Overview of the ribosome–EF-Tu complex. Left, cryo-EM map filtered to the local resolution of the factor binding site (3.1 Å). EF-Tu (red), the antibiotic kirromycin (blue), Phe–tRNA[Phe] in the A/T state (purple), fMet-tRNA[fMet] in the P site (green) and deacylated tRNA[fMet] in the E site (orange) are shown. Centre, comparison of the present structure with the X-ray structure of the kirromycin-stalled complex from *T. thermophilus* (*Tth*) with Trp–tRNA[Trp], PDB ID: 2Y18 (grey)[6]. Differences in tRNA structures are indicated by arrows, the dashed box shows the position of kirromycin. Right, comparison of experimental densities for kirromycin in this cryo-EM map (left) and crystallographic maps determined at 3.1 Å (centre, PDB ID: 2Y18[6]) and 3.5 Å resolutions (right, PDB ID: 2WRN[26]). The cryo-EM map filtered to the local resolution of about 3.4 Å is shown at 1.5σ (magenta mesh), the X-ray 2mF_o–DF_c maps at 1σ (blue mesh). b, Protein L9 on the ribosome. L9 in a bent

conformation with the carboxy-terminal domain (CTD) contacting ribosomal proteins (L9 cryo-EM, purple) versus the elongated L9 in crystal structures (L9 X-ray, pink, PDB ID: 4TP9[5]). In crystals, L9–CTD interacts with the 16S rRNA of a neighbouring ribosome rather than the 30S subunit of the same ribosome. c, Close up of L9–ribosome interactions marked by the dashed box in b. d, Coupling between L31 motions and global ribosome dynamics. Top, L31 in the non-rotated codon-recognition state of the ribosome with EF-Tu and three tRNAs bound. 5S, 5S rRNA; h42, helix 42 of 16S rRNA; NTD, N-terminal domain. Bottom, L31 in the rotated state of the ribosome with two tRNAs in hybrid states. L31 accommodates the large-scale rearrangements of the ribosome using its flexible linker region (black arrows), while preserving its binding sites on the ribosome. Mid left, schematic depicting the ratcheting of head (h) and body (b) of the 30S subunit.

of L9 observed in cryo-EM and crystals appear to be compatible with the binding of elongation factors to neighbouring ribosomes in polysomes as determined by cryo-electron-tomography[27] (Extended Data Fig. 8c). Furthermore, we built the model for L31, which is missing in the available *E. coli* 70S X-ray structures, probably owing to its dynamic nature. L31 bridges the ribosome at the top of the 30S head and the central protuberance of the 50S subunit. Upon 30S subunit ratcheting, the linker region of L31 switches from an extended to a kinked conformation, while L31 maintains its interactions with both subunits (Fig. 3d). We visualized this structural rearrangement by analysing another cryo-EM map of the ribosome with tRNAs in hybrid states (Extended Data Fig. 8b). The low occupancy of L31 can be explained by its flexible binding mode, which may also be important for the function of L31 as a Zn^{2+} reservoir for the cell[28].

In conclusion, our data shows that aberration-corrected cryo-EM allows dynamic macromolecular machines, such as the ribosome, to be visualized at a uniform resolution better than 3 Å with only small variations in local resolution. The 'purification' of electron microscopic images by computational sorting appears to be as powerful as the purification of ribosome conformation and composition during crystal growth in X-ray crystallography. The cryo-EM map visualizes bound water molecules, ions and rRNA modifications, providing novel insights into ribosome and antibiotic function, and thereby contributing to an improved structural basis for the development of new antibiotics.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

1. Leschziner, A. E. & Nogales, E. Visualizing flexibility at molecular resolution: analysis of heterogeneity in single-particle electron microscopy reconstructions. *Annu. Rev. Biophys. Biomol. Struct.* **36,** 43–62 (2007).
2. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
3. Yu, X., Ge, P., Jiang, J. S., Atanasov, I. & Zhou, Z. H. Atomic model of CPV reveals the mechanism used by this single-shelled virus to economically carry out functions conserved in multishelled reoviruses. *Structure* **19,** 652–661 (2011).
4. Wong, W. *et al.* Cryo-EM structure of the *Plasmodium falciparum* 80S ribosome bound to the anti-protozoan drug emetine. *eLife* **3,** e03080 (2014).
5. Noeske, J. *et al.* Synergy of streptogramin antibiotics occurs independently of their effects on translation. *Antimicrob. Agents Chemother.* **58,** 5269–5279 (2014).
6. Schmeing, T. M., Voorhees, R. M., Kelley, A. C. & Ramakrishnan, V. How mutations in tRNA distant from the anticodon affect the fidelity of decoding. *Nature Struct. Mol. Biol.* **18,** 432–436 (2011).
7. Fischer, N., Konevega, A. L., Wintermeyer, W., Rodnina, M. V. & Stark, H. Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature* **466,** 329–333 (2010).
8. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).
9. Polikanov, Y. S. *et al.* Amicoumacin A inhibits translation by stabilizing mRNA interaction with the ribosome. *Mol. Cell* **56,** 531–540 (2014).
10. Schmeing, T. M., Huang, K. S., Kitchen, D. E., Strobel, S. A. & Steitz, T. A. Structural insights into the roles of water and the 2′ hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol. Cell* **20,** 437–448 (2005).
11. Sergiev, P. *et al.* in *Ribosomes* (eds Rodnina, M. V., Wintermeyer, W. & Green, R.) Ch. 9, 97–110 (Springer Vienna, 2011).
12. Burakovsky, D. E. *et al.* Impact of methylations of m2G966/m5C967 in 16S rRNA on bacterial fitness and translation initiation. *Nucleic Acids Res.* **40,** 7885–7895 (2012).
13. Das, G. *et al.* Role of 16S ribosomal RNA methylations in translation initiation in *Escherichia coli. EMBO J.* **27,** 840–851 (2008).
14. Kimura, S. & Suzuki, T. Fine-tuning of the ribosomal decoding center by conserved methyl-modifications in the *Escherichia coli* 16S rRNA. *Nucleic Acids Res.* **38,** 1341–1352 (2010).
15. Schuwirth, B. S. *et al.* Structural analysis of kasugamycin inhibition of translation. *Nature Struct. Mol. Biol.* **13,** 879–886 (2006).
16. Gutierrez, B. *et al.* Fitness cost and interference of Arm/Rmt aminoglycoside resistance with the RsmF housekeeping methyltransferases. *Antimicrob. Agents Chemother.* **56,** 2335–2341 (2012).
17. Green, R. & Noller, H. F. *In vitro* complementation analysis localizes 23S rRNA posttranscriptional modifications that are required for *Escherichia coli* 50S ribosomal subunit assembly and function. *RNA* **2,** 1011–1021 (1996).
18. LaMarre, J. M., Howden, B. P. & Mankin, A. S. Inactivation of the indigenous methyltransferase RlmN in *Staphylococcus aureus* increases linezolid resistance. *Antimicrob. Agents Chemother.* **55,** 2989–2991 (2011).
19. Toh, S.-M. & Mankin, A. S. An indigenous posttranscriptional modification in the ribosomal peptidyl transferase center confers resistance to an array of protein synthesis inhibitors. *J. Mol. Biol.* **380,** 593–597 (2008).
20. Osterman, I. A. *et al.* Methylated 23S rRNA nucleotide m2G1835 of *Escherichia coli* ribosome facilitates subunit association. *Biochimie* **93,** 725–729 (2011).
21. Sergiev, P. V., Serebryakova, M. V., Bogdanov, A. A. & Dontsova, O. A. The ybiN gene of *Escherichia coli* encodes adenine-N6 methyltransferase specific for modification of A1618 of 23S ribosomal RNA, a methylated residue located close to the ribosomal exit tunnel. *J. Mol. Biol.* **375,** 291–300 (2008).
22. David-Eden, H., Mankin, A. S. & Mandel-Gutfreund, Y. Structural signatures of antibiotic binding sites on the ribosome. *Nucleic Acids Res.* **38,** 5982–5994 (2010).
23. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife* **1,** e00311 (2012).
24. Schröder, G. F., Levitt, M. & Brunger, A. T. Deformable elastic network refinement for low-resolution macromolecular crystallography. *Acta Crystallogr. D* **70,** 2241–2255 (2014).
25. Kleywegt, G. J. Crystallographic refinement of ligand complexes. *Acta Crystallogr. D* **63,** 94–100 (2007).
26. Schmeing, T. M. *et al.* The crystal structure of the ribosome bound to EF-Tu and aminoacyl-tRNA. *Science* **326,** 688–694 (2009).
27. Brandt, F. *et al.* The native 3D organization of bacterial polysomes. *Cell* **136,** 261–271 (2009).
28. Akanuma, G., Nanamiya, H., Natori, Y., Nomura, N. & Kawamura, F. Liberation of zinc-containing L31 (RpmE) from ribosomes by its paralogous gene product, YtiA, in *Bacillus subtilis. J. Bacteriol.* **188,** 2715–2720 (2006).
29. Samaha, R. R., Green, R. & Noller, H. F. A base pair between tRNA and 23S rRNA in the peptidyl transferase centre of the ribosome. *Nature* **377,** 309–314 (1995).
30. Ippolito, J. A. *et al.* Crystal structure of the oxazolidinone antibiotic linezolid bound to the 50S ribosomal subunit. *J. Med. Chem.* **51,** 3353–3356 (2008).

Author Contributions N.F. designed and performed cryo-EM experiments and data analysis. P.N. conceived and performed pseudo-crystallographic refinement and model validation and analysed data. A.L.K. prepared ribosome complexes. L.V.B. performed and analysed molecular dynamics simulations. All authors discussed the results. H.S. and N.F. conceived the project and wrote the paper with input from all authors.

Author Information The 2.9 Å cryo-EM map of the *E. coli* ribosome–EF-Tu complex has been deposited in the Electron Microscopy Data Bank with accession code EMD-2847, the coordinates of the atomic model have been deposited in the Protein Data Bank under accession code 5AFI. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.F. (niels.fischer@mpibpc.mpg.de) or H.S. (hstark1@gwdg.de).

METHODS

Complex preparation. To obtain kirromycin-stalled *E. coli* ribosome–EF-Tu complexes, ribosomes from *E. coli* MRE 600, initiation factors (IF1, IF2, IF3), fMet–tRNA^{fMet}, EF-Tu and Phe–tRNA^{Phe} were prepared as described[31–33]. Prior to initiation, the mRNA (GGCAAGGAGGUAAAUA*AUG*UU*C*GUUACGAC; the AUG start codon coding for fMet and UUC coding for Phe are underlined) was incubated with 0.1 mM EDTA for 90 s at 80 °C and shock cooled in an ice-water bath. 70S ribosomes (3 μM) were incubated with IF1, IF2, IF3 (4.5 μM), mRNA (15 μM), and f[^{3}H]Met–tRNA^{fMet} (7 μM) in buffer A (50 mM Tris-HCl, pH 7.5, 70 mM NH$_4$Cl, 30 mM KCl, 7 mM MgCl$_2$) containing 2 mM dithiothreitol (DTT) and 1 mM GTP for 30 min at 37 °C. Initiation efficiency was verified by nitrocellulose binding assay and radioactivity counting to be close to 100%. The complexes were purified by size-exclusion chromatography on a Biosuite 450 HR 5 μm column (Waters) using HPLC Alliance system (Waters). The cognate ternary complex EF-Tu–GTP–Phe–tRNA^{Phe} was prepared in buffer B (50 mM HEPES-KOH, pH 7.5, 70 mM NH$_4$Cl, 30 mM KCl, 7 mM MgCl$_2$, 2 mM DTT) using a twofold excess of EF-Tu over Phe–tRNA^{Phe}. Initiation complexes (0.07 μM) were mixed with excess of deacylated tRNA^{fMet} (0.2 μM) and ternary complexes (0.12 μM) in buffer C (50 mM HEPES-KOH, pH 7.5, 70 mM NH$_4$Cl, 30 mM KCl, 20 mM MgCl$_2$, 1 mM DTT, 0.6 mM spermine, 0.4 mM spermidine) in the presence of 150 μM kirromycin.

Cryo-EM analysis. Five microlitres of 0.07 μM kirromycin-stalled ribosome-EF-Tu complexes were applied onto cryo-EM grids (Quantifoil 3.5/1 μm, Jena) covered with pre-floated continuous carbon. Cryo-EM grids were prepared under controlled environmental conditions at 4 °C and 100% humidity in a vitrification device (Vitrobot Mark IV, FEI Company, Eindhoven). A total of 24,684 image frames were recorded in spot-scanning mode (4 × 4 image frames per 3.5 μm hole in the Quantifoil carbon film) on a Falcon I direct detector under liquid-nitrogen conditions with a Titan Krios electron microscope (FEI, Eindhoven) equipped with a XFEG electron source and a C$_s$ corrector (CEOS, Heidelberg) using 300 kV acceleration voltage, an electron dose of ~40 ± 5 electrons per Å^2 (determination based on calibration with a Faraday cup), −0.7 to −2.5 μm defocus and a nominal magnification of 192,000×, resulting in a final pixel size of ~0.755 Å (as determined by comparing the final 3D map with atomic models from X-ray crystallography). Using the C$_s$ corrector, electron optical aberrations were corrected to residual phase errors of 45° at scattering angles of >12 to 15 mrad (that is, less than 45° phase error at 1.8 to 2.1 Å; Extended Data Fig. 1). Linear geometrical distortions were reduced to <0.1% using the *tilthexapole* beam coils of the C$_s$ corrector. Furthermore, coma caused by the spot-scanning procedure was minimized using the *usrimageshift* correction in the C$_s$ corrector alignment. Ribosome particle images were extracted in a fully automated manner using template-independent custom-made software (CowPicker, B. Busche and H.S., unpublished data). The 1,603,254 extracted particle images were corrected locally for the contrast-transfer function by classification and averaging and selected according to quality of powerspectra[34], that is, to show Thon rings better than 3.4 Å up to 2.4 Å. In all subsequent steps, the resulting 1,339,775 contrast-transfer-function-corrected particle images were used. First, the particle images were sorted into groups of particles according to: (1) 30S body rotation, as described[7]; and (2) ligand occupancy, using supervised classification by projection matching[35] on the basis of a structural library of different ribosome complexes[7,36] (Extended Data Fig. 2). Sorting in both steps was performed using low-pass-filtered reference maps and particle images binned to about 6 Å per pixel. Finally, 3D classification in RELION 1.2[37] was used to obtain the final set of 417,201 particle images with good elongation factor for the refinement to high-resolution (2.9 Å) using the 'gold-standard procedure' in RELION 1.2[37] (Extended Data Fig. 3). A local resolution map computed from the two unprocessed half-maps by Resmap[2] revealed only few variations in local resolution over the entire ribosome complex (Fig. 1a). We obtained another cryo-EM map of a ratcheted ribosome with tRNAs in hybrid states and showing protein L31 in a distinct conformation by sorting particle images with 10° rotation angles according to ligand occupancy, as described above, resulting in a final homogeneous population of 8,073 particles, which was refined to 6.4 Å resolution (0.143 criterion) using the gold-standard procedure in RELION.

Atomic fluctuations obtained from molecular dynamics simulations. The molecular dynamics simulation of the ribosome in a pre-translocation state (pre1a) presented earlier[38] was extended to 2 μs. To identify the rigid core of the 50S subunit, the root mean square fluctuation (r.m.s.f.) of each atom was calculated using the program g_rmsf from the GROMACS simulation suite[39] after alignment to the 50S subunit, omitting the first 0.5 μs. The rigid core was defined as all 50S atoms excluding those of the tail of the r.m.s.f. histogram starting at the point where the frequency drops to half of the maximum frequency (0.19 nm). Finally, the r.m.s.f. of all atoms after alignment to the rigid core was calculated (Extended Data Fig. 6).

Pseudo-crystallographic refinement and model building. For initial model building, the cryo-EM density map was sharpened by applying a B factor of −120 Å^2, filtered to 3.1 Å resolution and masked using a pseudo bulk solvent envelope obtained by merging different versions of the cryo-EM map filtered at different frequencies (12.0 Å, 8.0 Å, 6.0 Å and 4.0 Å) with the RAVE package[40] and Chimera[41]. Fourier transform of the masked cryo-EM density map to reciprocal space structure factors was performed using Crystallography and NMR System (CNS)[42,43] employing phase significance blurring scale factors derived from FSC values in a resolution-dependent manner[44]. In detail, a modified CNS input file was used for the assignment of FOMs (figure of merit) estimated based on equation (1) for the map obtained from the full data set and equation (2) for the maps obtained from two half sets:

$$FOM = \left(\frac{2 \times FSC}{(1 + FSC)} \right)^{0.5} \qquad (1)$$

$$FOM = (FSC)^{0.5} \qquad (2)$$

Obtained phase probabilities, written in the form of Hendrickson–Lattman (HL) coefficients, were used for reciprocal space refinement, performed against an MLHL target (maximum likelihood with experimental phase probability distribution) in both CNS and PHENIX[45] programs using both X-ray and electron scattering factors, respectively. Both programs employed automatic optimization of weights used to balance the relative contributions of experimental and restraints terms using a grid search. To fulfil the requirements of the crystallographic MLHL refinement, 5% of the reflections were selected randomly for the 'Rfree' set, which was kept identical for all refinements. Homology modelling combined with density-guided energy optimization was performed using the Rosetta package[46,47] employing templates and alignments provided by the HHPRED server[48]. Model density maps were generated based on finally refined models, without bulk solvent correction as implemented in PHENIX. Map normalization (mean and standard deviation of density values are 0 and 1.0, respectively) was performed using MAPMAN (Rave package).

The initial fit of an atomic model of the *E.coli* ribosome assembled from various crystal structures (PDB codes: 4GD2[49] (30S)[49], 3R8T (50S)[49], 2J00 (mRNA, tRNA^{fMet})[50], 3L0U (tRNA^{Phe})[51], 1OB2 (EF-Tu) (R. C. Nielsen *et al.* unpublished data)) was performed using Chimera, followed by rigid body refinement in the PHENIX program. The atomic model was refined with deformable elastic network (DEN) restraints[52] in CNS with alternating cycles of manual rebuilding in Coot[53] and monitoring the local fit to the density with RESOLVE[54]. In addition the overall refinement progress was monitored by calculating CC$_{work}$ and CC$_{free}$ (ref. 8), as well as the correlation between the cryo-EM and the model map (FSC$_{work}$). The one-dimensional structure factor derived by rotational averaging from the initially refined model was used to re-sharpen the raw cryo-EM density map. Prior to Fourier transformation, the re-sharpened, normalized 3D cryo-EM map was solvent flattened using a smoothed model-based envelope (Rave package, Chimera), encompassing the volume within the distance of at least 3.0 Å from each atom. As the density in the protein region is not supposed to be negative, voxels with negative density in the protein region were set to zero during the solvent flattening process by the MAPMASK program in the CCP4 suite[55]. Further model improvement and fitting were facilitated by real space refinement (ERRASER[56] and phenix.real_space_refine[57] against the map calculated using the working set of reflections only) and manual corrections in Coot combined with reciprocal space refinement. Modelling of post-transcriptional modifications was performed in Coot and was based on thorough analysis of the cryo-EM map. Modifications resulting from addition or substitution of an atom or atoms in comparison with unmodified bases were modelled if the presence of additional atoms or consequent changes in shape (for dihydrouridine) were supported by the cryo-EM map. Pseudouridines were modelled as indicated by additional polar/hydrogen-bond interactions formed by the additional amine group. For one methylation (m^{5}747 in 23S rRNA) showing no well-defined corresponding density and two pseudouridines (Ψ746 and Ψ2457 in 23S rRNA) with no clear additional interactions, modifications were modelled on the basis of ref. 11 and references therein. In all other cases manual modelling of modified nucleotides was performed only if the cryo-EM map was well defined. New stereochemistry definitions for non-standard ligands were generated with phenix.reel. A homology model of protein L31 was built in Rosetta using the *T. thermophilus* X-ray structure (PDB ID: 3I8I)[58] and constraints from the present cryo-EM map filtered to 4.5 Å resolution. Models of protein L9 and L31 were manually adjusted to fit the cryo-EM map filtered at 3.9 Å and 4.5 Å resolution, respectively in the programs O (ref. 59) and Coot. The models were further refined in real space using Rosetta followed by phenix.real_space_refine against the map created from working set reflections only. To maintain the intermolecular interactions of L9 and L31 proteins, the atomic models used for real space refinement in phenix.real_space_refine included surrounding proteins and parts of RNA chains within a radius of at least 10 Å. Improved protein models were included in the overall model used for the final reciprocal space refinement against 2.9 Å resolution data generated from the sharpened and solvent-flattened 3D cryo-EM map in PHENIX. The final model consisting of 152,718 individual atoms has been refined to 24.08% and 0.922 for R$_{work}$ and CC$_{work}$ (definition is given below), respectively. The final model exhibits a good stereochemistry

with 85.59% of residues in the most favoured region and 3.30% residues in the disallowed region of the Ramachandran plot, protein side chain outliers of 2.63% and all atom clash score 9.15. Detailed refinement statistics are presented in Extended Data Table 1. The model for another conformer of protein L31 found in a cryo-EM map of the ribosome in complex with two tRNAs in hybrid states was built at 6.5 Å resolution, as described above for L9 and L31.

Refinement of half maps. Refinement of the final model against data sets obtained from two half maps was performed at a resolution of 2.9 Å in PHENIX using five to seven cycles of combined positional (real and reciprocal space) and atomic displacement parameter (ADP) refinement combined with automatically identified TLS (translation/libration/screw) groups. To remove possible model bias from the model refined against reflections obtained from the reconstruction using all the particles, a similar strategy was used as for higher-resolution refinements (a 0.5 Å random shift and additional restraints), as described below. The FSC and CC were calculated between the model and the half map used for refinement, as well as between the model and the other half map for cross-validation (Extended Data Fig. 3b).

Higher-resolution refinements. The final atomic model of the *E.coli* ribosome was divided into three sub-models containing about 65%, 43% and 24% of all residues, respectively, with each sub-model exhibiting similar B factors and resolution estimates for the corresponding fragment of the cryo-EM map (Extended Data Fig. 3). The resulting three models were used to calculate smoothed masks encompassing the volume within a distance of at least 2.5 Å from each atom. Those masks were used to cut the required portion of the cryo-EM density map using a pseudo bulk solvent flattening procedure before conversion to reciprocal space structure factors, as described above. Phase significance blurring scale factors applied during the Fourier transform were derived from FSC values calculated between fragments of the two half maps masked by model-based envelopes. The refinement of individual models (initially to 2.75 Å, 2.60 Å and 2.50 Å for 65%, 43% and 24% of all residues, respectively) against MLHL targets were performed in PHENIX by gradually decreasing the high-resolution limit in 0.05 Å steps, for example, local resolution refinements starting with 2.50 Å were performed at 2.50 Å, 2.55 Å, 2.60 Å, 2.65 Å, 2.70 Å and 2.75 Å. To eliminate possible bias against the model refined at 2.9 Å resolution, the atoms were displaced by a random translation up to 0.1 Å before performing the full refinement using restraints for secondary structure, base-pairing, base-planarity and hydrogen bonding. The pseudo crystallographic high-resolution limit was selected based on several criteria: (1) the overall R and R_{work} factors (lower than 1/10 of the highest resolution limit); (2) the Pearson correlation coefficient calculated between F_{model} and F_{EM} (CC_{work}) used for refinement should be greater than 0.2 for the highest resolution shell and the overall correlation coefficient ($CC_{overall}$) should not be lower than 0.9; (3) the calculated FSC_{work} value between model map coefficients (F_{model}, $phase_{model}$) and structure factors derived from the cryo-EM map (F_{EM}, $phase_{EM}$) used for refinement should be greater than 0.5 for the highest resolution shell; (4) the calculated crystallographic R factor for the highest resolution shell should be not greater than 51%. The mentioned statistical values have been calculated in a resolution-dependent manner using PHENIX and SFALL (CCP4 suite) for 20 shells comprising a similar number of reflections (default number of shells for reporting refinement statistics in PHENIX).

Preparation of figures. Figures were generated using Pymol 1.7 (Schrödinger, LLC.), UCSF CHIMERA 1.9[41], AMIRA 5.2.2 (Visage Imaging, Richmond VIC) and PRISM 5 (GraphPad Software, La Jolla).

31. Milon, P. *et al.* Transient kinetics, fluorescence, and FRET in studies of initiation of translation in bacteria. *Methods Enzymol.* **430,** 1–30 (2007).
32. Rodnina, M. V. *et al.* Thiostrepton inhibits the turnover but not the GTPase of elongation factor G on the ribosome. *Proc. Natl Acad. Sci. USA* **96,** 9586–9590 (1999).
33. Rodnina, M. V. & Wintermeyer, W. GTP consumption of elongation factor Tu during translation of heteropolymeric messenger-RNAs. *Proc. Natl Acad. Sci. USA* **92,** 1945–1949 (1995).
34. Sander, B., Golas, M. M. & Stark, H. Automatic CTF correction for single particles based upon multivariate statistical analysis of individual power spectra. *J. Struct. Biol.* **142,** 392–401 (2003).
35. Valle, M. *et al.* Cryo-EM reveals an active role for aminoacyl-tRNA in the accommodation process. *EMBO J.* **21,** 3557–3567 (2002).
36. Stark, H. *et al.* Ribosome interactions of aminoacyl-tRNA and elongation factor Tu in the codon-recognition complex. *Nature Struct. Mol. Biol.* **9,** 849–854 (2002).
37. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
38. Bock, L. V. *et al.* Energy barriers and driving forces in tRNA translocation through the ribosome. *Nature Struct. Mol. Biol.* **20,** 1390–1396 (2013).
39. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29,** 845–854 (2013).
40. Kleywegt, G. J. & Jones, T. A. Software for handling macromolecular envelopes. *Acta Crystallogr. D* **55,** 941–944 (1999).
41. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
42. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2,** 2728–2733 (2007).
43. Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
44. Greber, B. J. *et al.* The complete structure of the large subunit of the mammalian mitochondrial ribosome. *Nature* **515,** 283–286 (2014).
45. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221 (2010).
46. DiMaio, F. *et al.* Improved molecular replacement by density-and energy-guided protein structure optimization. *Nature* **473,** 540–543 (2011).
47. DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. Refinement of protein structures into low-resolution density maps using rosetta. *J. Mol. Biol.* **392,** 181–190 (2009).
48. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–W248 (2005).
49. Dunkle, J. A. *et al.* Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* **332,** 981–984 (2011).
50. Selmer, M. *et al.* Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **313,** 1935–1942 (2006).
51. Byrne, R. T., Konevega, A. L., Rodnina, M. V. & Antson, A. A. The crystal structure of unmodified tRNA^{Phe} from *Escherichia coli*. *Nucleic Acids Res.* **38,** 4154–4162 (2010).
52. Schröder, G. F., Levitt, M. & Brunger, A. T. Super-resolution biomolecular crystallography with low-resolution data. *Nature* **464,** 1218–1222 (2010).
53. Emsley, P., Lohkamp, B., Scott, W. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
54. Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56,** 965–972 (2000).
55. The CCP4 Suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
56. Chou, F.-C., Sripakdeevong, P., Dibrov, S. M., Hermann, T. & Das, R. Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nature Methods* **10,** 74–76 (2013).
57. Afonine, P. V., Headd, J. J., Terwilliger, T. C. & Adams, P. D. New tool: phenix.real_space_refine. *Computational Crystallography Newsletter* **4,** 43–44 (2013).
58. Jenner, L., Demeshkina, N., Yusupova, G. & Yusupov, M. Structural rearrangements of the ribosome at the tRNA proofreading step. *Nature Struct. Mol. Biol.* **17,** 1072–1078 (2010).
59. Jones, T. A. & Kjeldgaard, M. Electron-density map interpretation. *Methods Enzymol.* **277,** 173–208 (1997).
60. Allegretti, M., Mills, D. J., McMullan, G., Kühlbrandt, W. & Vonck, J. Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. *eLife* **3,** e01963 (2014).
61. Bartesaghi, A., Matthies, D., Banerjee, S., Merk, A. & Subramaniam, S. Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *Proc. Natl Acad. Sci. USA* **111,** 11709–11714 (2014).
62. Burmeister, W. P. Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Crystallogr. D* **56,** 328–341 (2000).
63. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333,** 721–745 (2003).
64. Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework. *J. Appl. Crystallogr.* **35,** 126–136 (2002).
65. Demirci, H. *et al.* Modification of 16S ribosomal RNA by the KsgA methyltransferase restructures the 30S subunit to optimize ribosome function. *RNA* **16,** 2319–2324 (2010).
66. Helser, T. L., Davies, J. E. & Dahlberg, J. E. Mechanism of kasugamycin resistance in *Escherichia coli*. *Nature* **235,** 6–9 (1972).
67. O'Connor, M., Thomas, C. L., Zimmermann, R. A. & Dahlberg, A. E. Decoding fidelity at the ribosomal A and P sites: influence of mutations in three different regions of the decoding domain in 16S rRNA. *Nucleic Acids Res.* **25,** 1185–1193 (1997).
68. François, B. *et al.* Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding. *Nucleic Acids Res.* **33,** 5677–5690 (2005).
69. Jiang, J., Aduri, R., Chow, C. S. & SantaLucia, J. Structure modulation of helix 69 from *Escherichia coli* 23S ribosomal RNA by pseudouridylations. *Nucleic Acids Res.* **42,** 3971–3981 (2013).
70. Davis, D. R. Stabilization of RNA stacking by pseudouridine. *Nucleic Acids Res.* **23,** 5020–5026 (1995).
71. Ortiz-Meoz, R. F. & Green, R. Helix 69 is key for uniformity during substrate selection on the ribosome. *J. Biol. Chem.* **286,** 25604–25610 (2011).
72. Seidelt, B. *et al.* Structural insight into nascent polypeptide chain–mediated translational stalling. *Science* **326,** 1412–1415 (2009).
73. Pulk, A. & Cate, J. H. Control of ribosomal subunit rotation by elongation factor G. *Science* **340,** 1235970 (2013).
74. Vorstenbosch, E., Pape, T., Rodnina, M., Kraal, B. & Wintermeyer, W. The G222D mutation in elongation factor Tu inhibits the codon-induced conformational changes leading to GTPase activation on the ribosome. *EMBO J.* **15,** 6766–6774 (1996).
75. Daviter, T., Wieden, H.-J. & Rodnina, M. V. Essential role of histidine 84 in elongation factor Tu for the chemical step of GTP hydrolysis on the ribosome. *J. Mol. Biol.* **332,** 689–699 (2003).

76. Voorhees, R. M., Schmeing, T. M., Kelley, A. C. & Ramakrishnan, V. The mechanism for activation of GTP hydrolysis on the ribosome. *Science* **330,** 835–838 (2010).
77. Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* **170,** 427–438 (2010).
78. Selmer, M., Gao, Y.-G., Weixlbaumer, A. & Ramakrishnan, V. Ribosome engineering to promote new crystal forms. *Acta Crystallogr. D* **68,** 578–583 (2012).

Hydrogens detected by subatomic resolution protein crystallography in a [NiFe] hydrogenase

Hideaki Ogata[1], Koji Nishikawa[1]† & Wolfgang Lubitz[1]

The enzyme hydrogenase reversibly converts dihydrogen to protons and electrons at a metal catalyst[1]. The location of the abundant hydrogens is of key importance for understanding structure and function of the protein[2–6]. However, in protein X-ray crystallography the detection of hydrogen atoms is one of the major problems, since they display only weak contributions to diffraction and the quality of the single crystals is often insufficient to obtain sub-ångström resolution[7]. Here we report the crystal structure of a standard [NiFe] hydrogenase (~91.3 kDa molecular mass) at 0.89 Å resolution. The strictly anoxically isolated hydrogenase has been obtained in a specific spectroscopic state, the active reduced Ni-R (subform Ni-R1) state. The high resolution, proper refinement strategy and careful modelling allow the positioning of a large part of the hydrogen atoms in the structure. This has led to the direct detection of the products of the heterolytic splitting of dihydrogen into a hydride (H^-) bridging the Ni and Fe and a proton (H^+) attached to the sulphur of a cysteine ligand. The Ni–H^- and Fe–H^- bond lengths are 1.58 Å and 1.78 Å, respectively. Furthermore, we can assign the Fe–CO and Fe–CN^- ligands at the active site, and can obtain the hydrogen-bond networks and the preferred proton transfer pathway in the hydrogenase. Our results demonstrate the precise comprehensive information available from ultra-high-resolution structures of proteins as an alternative to neutron diffraction and other methods such as NMR structural analysis.

[NiFe] hydrogenases are composed of two main subunits (molecular masses ~28.8 and 62.5 kDa) that harbour several metal centres and provide pathways for dihydrogen, protons and electrons during the catalytic reaction of the enzyme (Fig. 1)[1–3]. The active site located in the centre of the molecule is composed of a dinuclear Ni–Fe complex. The iron is coordinated by one CO and two CN^- ligands. Two thiolates of cysteine residues are bridging the metals. Furthermore, the Ni ion is coordinated by two terminal cysteine thiolates. The third bridging ligand between the Ni and Fe atoms changes identity in the catalytic cycle, in which the enzyme passes through several intermediate redox states (see Fig. 1b)[4–7]. In the inactive oxidized form, the third bridging ligand is OH^-. In the active form, three different redox states are known as Ni-SI$_a$, Ni-C and Ni-R (Fig. 1b). Once the enzyme is activated by hydrogen reduction, the bridging ligand (OH^-) is removed as revealed by a previous X-ray crystal structure at 1.4 Å resolution[8]. One-electron reduction of the enzyme yields the paramagnetic Ni-C state[8–10]. This state carries a hydride bridge between the Ni and Fe as shown by electron paramagnetic resonance (EPR) techniques[9–11]. Upon illumination at cryogenic temperature this bridge is lost and the associated proton moves to a nearby base (Ni-L state)[9,12]. One electron reduction of Ni-C leads to the most reduced Ni-R state, which exists in three isoelectronic subforms (Ni-R1, Ni-R2 and Ni-R3) distinguishable by Fourier transform infrared (FTIR) spectroscopy[13]. The detailed structure of the Ni-R states is not yet known, but is of key importance for understanding the catalytic mechanism. We therefore performed an X-ray crystallographic analysis of the Ni-R state of [NiFe] hydrogenase from *Desulfovibrio vulgaris* Miyazaki F at 0.89 Å resolution. The ultra-high resolution led to a very detailed geometric structure of the [NiFe] hydrogenase, including the

active site, Mg ion and iron–sulphur clusters (Fig. 2 and Extended Data Fig. 1). In particular, a large number of hydrogens could be assigned, even near the catalytically active [NiFe] site, where a hydride bridge and a protonated cysteine residue could be detected for the Ni-R state, formed in the first step of heterolytic dihydrogen splitting by the enzyme.

Inactivation or oxidative damage of the enzyme was avoided by performing isolation, purification and crystallization of the hydrogenase under strict anoxic conditions. FTIR spectroscopy was used to monitor the redox state of the enzyme in the resulting crystals (under 1.5% H_2) and showed that they essentially contained a single catalytic intermediate, the fully reduced Ni-R state (see Extended Data Fig. 2).

Detection of hydrogen atoms near the metal(s) by X-ray crystallography is difficult owing to the Fourier truncation ripples that appear around the heavy atoms in the electron density map. It is known that the occurrence of these ripples depends on the resolution and the atomic displacement parameters (B-factors)[14,15]. Therefore, we analysed two data sets with different resolution. The crystal structures (termed Ni-R$_{UH}$ and Ni-R$_H$) were determined at 0.89 Å and 1.06 Å resolution, respectively. The structure refinement was performed with the PHENIX program suite[16]. The final R/R_{free} factors converged to 9.6/10.6% for Ni-R$_{UH}$ and 11.0/13.2% for Ni-R$_H$ (Extended Data Table 1). No redox change of the [NiFe] active site by the X-ray exposure and no radiation damage (for
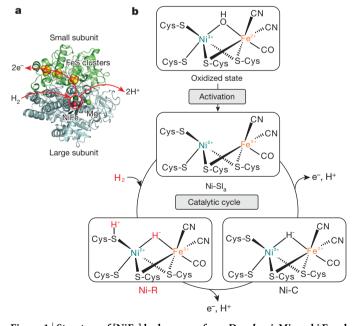

Figure 1 | Structure of [NiFe] hydrogenase from *D. vulgaris* Miyazaki F and proposed catalytic cycle. a, Overall structure of [NiFe] hydrogenase protein. Colour code of cofactors: nickel, green; iron, orange; magnesium, light green; carbon, grey; nitrogen, blue; oxygen, red; sulphur, yellow. **b**, Proposed catalytic reaction mechanism of [NiFe] hydrogenase; the different states are named according to the active metal site (Ni–X)[1,28].

[1]Max Planck Institute for Chemical Energy Conversion, Stiftstrasse 34-36, D-45470 Mülheim an der Ruhr, Germany. †Present address: Department of Life Science, Graduate School of Life Science, University of Hyogo, 3-2-1 Koto, Kamigori-cho, Ako-gun, Hyogo 678-1297, Japan.
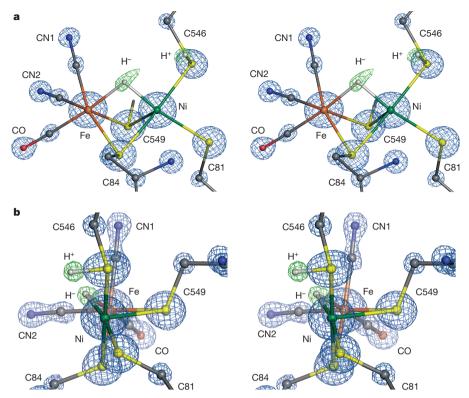
Figure 2 | The [NiFe] active site in the Ni-R form. a, Stereo view of the $2F_o - F_c$ electron density map contoured at 4σ and the $F_o - F_c$ omit map contoured at 5.5σ of the Ni-R_{UH} data shown in blue and green, respectively.

The green mesh indicates the electron density of the hydrogen atoms. **b**, Stereo view of the $2F_o - F_c$ electron density map contoured at 3σ and the $F_o - F_c$ omit map contoured at 6σ shown as a side view of the Ni-R_H data.

example, decarboxylation of Glu/Asp, and S–C bond cleavage of Met) were observed. The bond length estimated standard deviations by least-square full-matrix inversion were 0.002–0.007 Å for the [NiFe] active site and >0.006 Å for the amino-acid residues, respectively (Extended Data Fig. 3a). For the hydrogen atoms, the residual electron densities in the $F_o - F_c$ omit map are slightly shifted towards the attached heavier atoms (X) in the X-ray structure compared with the neutron diffraction structure, which directly observes the nuclei[17]. Furthermore, we observed that the peaks of the residual electron densities are slightly off position from the riding models (Extended Data Fig. 3b). The coordinate errors of the X–H bond distances for the amino-acid residues were estimated to <0.2 Å. For the Ni-R_{UH} structure, 93% of the hydrogen atoms on the selected 50 amino-acid residues located within a 10 Å radius around the Ni could be confirmed in the $F_o - F_c$ electron density omit map (2σ contour level). The hydrogen atoms of 124 water molecules in the interior of the protein were also identified in this $F_o - F_c$ omit map. The subatomic resolution structure allowed us to determine extended hydrogen bond networks involving amino acids and water molecules and the precise water coordination of the Mg^{2+} ion (metal-aquo centre).

We confirmed the location of the hydride bridge between Ni and Fe in the active site of Ni-R (Fig. 2 and Extended Data Figs 4 and 5). A further detail is that the Ni–H^- distance seems to be shorter than the Fe–H^- distance (Table 1). For the assignment of the bridging ligand, we excluded two possibilities: (1) partial occupation by heavier anions (OH^- or SH^-) and (2) contamination by oxidized states in the sample. As shown in Extended Data Table 2 the observed distance (1.57 Å) between Ni and the residual electron density peak is much shorter than the distances reported for Ni^{2+}–OH^- or Ni^{2+}–SH^-, which are 1.9 and 2.3 Å, respectively. The observed bond distance (estimated standard deviation 0.2 Å) is in agreement with distances obtained from Ni–hydride complexes determined by neutron diffraction (Extended Data Table 2). Furthermore, the oxygen species found in the oxidized states did not fit in the residual electron density (see Extended Data Fig. 6). The Ni–Fe distance was obtained to 2.57 Å establishing a weak metal–metal bond[18].

For the related paramagnetic Ni-C state (Ni^{3+}) the analysis of magnetic resonance data (HYSCORE, hyperfine sublevel correlation spectroscopy) also showed an asymmetric hydride bridge (Ni–H^- distances of 1.59 Å)[9]. The hydrogenase model complexes for Ni-R reported so far show the bridging hydride rather bound to the Fe^{2+} than to the Ni^{2+} atom, which is different from the hydrogenase[19,20].

The Ni-R and Ni-SI_a states are EPR-silent with a divalent Ni^{2+} (Fig. 1b), which could be either low-spin or high-spin. The latter should be favoured by a tetrahedral ligand sphere, the former by a square-planar ligand arrangement. In the Ni-R_{UH} structure, the distances of the Ni ion to the thiolates of the cysteine residues were 2.2–2.3 Å except for Ni–S(Cys 549), which has a distance of >2.5 Å (Table 1 and Extended Data Fig. 7). This weak axial ligation makes the nickel-ion ligand sphere essentially square pyramidal, indicating that the Ni^{2+} is probably in a low-spin state.

After heterolytically splitting hydrogen (in the Ni-SI_a to Ni-R transition; Fig. 1b), a proton is expected to be transferred from the active site to the molecular surface. One of the cysteine sulphurs coordinating the nickel is expected to function as the initial proton acceptor[12]. In the present structure the residual electron density at the thiolate of Cys 546 is assigned to a protonated state of the cysteine (Fig. 2). The distance between the sulphur and the proton is 1.1–1.3 Å, in agreement with the expected theoretical values (1.2 Å), although the angle (C–S–H) is slightly smaller than the expected one (Table 1). Moreover, the electron density of the thiolate of Cys 546 shows a somewhat higher thermal motion than the other ligated thiolates, shown by the ellipsoid shape of the electron density (Fig. 2b and Extended Data Fig. 4). The thermal motion perpendicular to the Ni–S bond direction indicates reactivity of the thiol ligand as a proton acceptor, although the interatomic distances of Ni–S(Cys 546) did not change during the activation cycle. The metal–hydride distances might be sensitive to the protonation of the terminal thiolates. Ni^{2+}–S(H)R would be more Lewis-acidic than Ni^{2+}–SR. The distance of Ni–H^-···H^+–S(Cys 546) was 2.45 Å. Other protons near to the hydride were found on C_β of Cys 546 (2.59 Å) and on C_δ of Arg 479

Table 1 | Geometry of the [NiFe] active site*

State	Ni-R	Ni-R	Ni-R	Ni-C	Ni-C	Ni-C
Methods	X-ray†	X-ray‡	DFT§	DFT‖	DFT¶	QM/MM#
Ni–Fe	2.57	2.58	2.55	2.55	2.57	2.58
Ni–H⁻	1.58	1.57	1.58	1.60	1.59	
Fe–H⁻	1.78	1.72	1.72	1.72	1.70	
Fe–CO	1.75	1.73	1.73	1.73	1.73	1.74
Fe–CN1	1.88	1.87	1.86	1.86	1.89	1.87
Fe–CN2	1.91	1.89	1.86	1.87	1.90	1.87
C–O	1.15	1.18	1.17	1.17	1.18	1.18
C–N1	1.17	1.15	1.18	1.18	1.19	1.19
C–N2	1.15	1.19	1.18	1.18	1.19	1.19
Ni–S(Cys 81)	2.24	2.23	2.24	2.25	2.25	2.23
Ni–S(Cys 84)	2.21	2.22	2.19	2.27	2.31	2.25
Ni–S(Cys 546)	2.18	2.17	2.20	2.18	2.22	2.20
Ni–S(Cys 549)	2.54	2.55	2.55	2.35	2.35	2.33
Fe–S(Cys 84)	2.26	2.27	2.27	2.31	2.31	2.30
Fe–S(Cys 549)	2.31	2.32	2.37	2.32	2.34	2.33
S(Cys 81)–Ni–H⁻	167.5	167.1	166.5	167.8	168.0	
S(Cys 84)–Ni–S(Cys 546)	176.3	176.7	173.8	167.7	144.8	
S(Cys 546)–H★	1.09	1.31	1.36	—	—	—
C–S(Cys 546)–H	72.4	74.7	96.8	—	—	—
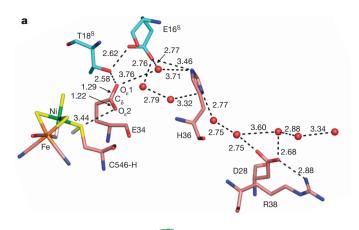
* Bond lengths and angles are shown in ångströms and degrees, respectively.
† Ni-R$_{UH}$ data in this work.
‡ Ni-R$_{H}$ data in this work.
§ DFT calculations with a hydride bridge and a protonated cysteine (Cys 546) in the Ni-R state from *D. vulgaris* Miyazaki F [NiFe] hydrogenase[28].
‖ DFT calculations with a hydride bridge in the Ni-C state from *D. vulgaris* Miyazaki F [NiFe] hydrogenase[28].
¶ DFT calculations with the fully relaxed structural optimization from *D. vulgaris* Miyazaki F [NiFe] hydrogenase[29].
# Quantum mechanics/molecular mechanics (QM/MM) calculations based on the membrane-bound [NiFe] hydrogenase from *Ralstonia eutropha*[30].
★ The peak of the electron density for the terminal-bound proton appears slightly shifted towards the attached heavy atom (sulphur in this case). The ideal S–H bond distance is 1.2 Å in X-ray crystallography, and 1.3 Å in case of the neutron diffraction as well as in DFT calculation.

(2.43 Å), respectively. These are not acidic and will not be involved in the catalytic reaction. No water molecules were found in the vicinity of the hydride bridge. In a recently described model system of the [NiFe] site, the protonation at a terminal thiolate ligand at the Ni could be demonstrated[21].

The postulated proton transfer pathway in the [NiFe] hydrogenase suggests that the proton is transferred from the initial cysteine acceptor (Cys 546) to a nearby glutamic acid (Fig. 3a)[22]. The protonation state of this Glu 34 can be obtained from the bond length of the carboxylic group, showing differences between C–OH (1.29 Å) and C=O (1.22 Å) (Fig. 3a). Glu 34 has a very short hydrogen bond of 2.58 Å to Thr 18^S (S indicates the small subunit) which has another low-barrier hydrogen bond[23] to Glu 16^S (2.62 Å), indicating their possible involvement in the proton transfer. For the further proton transfer, there are three major pathways (Fig. 3a and Extended Data Fig. 8)[24,25]. The two pathways, path 2 and path 3 (Extended Data Fig. 8a), are less favourable owing to the participation of non-conserved amino acids[26]. In path 1 well conserved amino-acid residues and water molecules are placed in similar positions as in other hydrogenases (Extended Data Fig. 8d). In the present structure His 36, which is located in the centre of path 1, was also observed in a protonated state (Extended Data Fig. 8b), indicating that it is participating in the proton transfer pathway.

The comparison of the X-ray crystallographic structures of the inactive oxidized state (at 1.04 Å resolution; Protein Data Bank (PDB) accession number 1WUI)[27] and the catalytically active Ni-R state in this work shows that upon H_2 reduction of the enzyme no significant conformational changes occur (root mean squared deviation = 0.089 Å). Moreover, the locations of the water molecules in the enzyme are almost identical in both states.

FTIR spectroscopic studies have shown that one CO and two CN⁻ are ligated to the low-spin Fe^{2+} atom in the [NiFe] active site[13]. The CO and CN⁻ ligands are isoelectronic so that it is difficult to distinguish them based on their electron density. However, the bond distances between the Fe and the carbon atoms show significant differences in the present structure, i.e. Fe–CO 1.75 Å and Fe–CN⁻ 1.88/1.91 Å (Table 1).
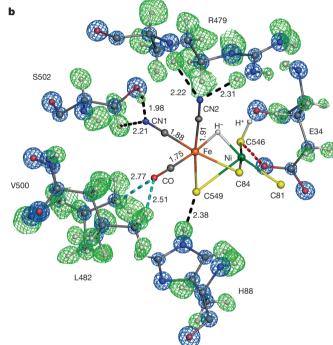

Figure 3 | A possible proton transfer pathway and the ligands of the [NiFe] active site. a, Stick representation of a possible proton transfer pathway (Path 1). The other possible proton pathways (Path 2 and Path 3) are shown in Extended Data Fig. 8. The superscript S indicates the amino-acid residues in the small subunit. **b**, Electron density maps of the amino-acid residues that are hydrogen bonded to the active site. The $2F_o − F_c$ map is contoured at 4σ and the $F_o − F_c$ omit map at 3.5σ of Ni-R$_{UH}$. The green mesh indicates the electron density of the hydrogen atoms. Hydrogen bonds to the cyanides and the bridging Cys 549 are shown as black dashes. The light blue dashes show the distances between the CO ligands and the hydrogens of the amino acids. The red dotted line shows a possible proton transfer path. The numbers shown in **a** and **b** represent the bond distances (in ångströms).

The CN⁻ ligands form two hydrogen bonds each with CN···H distances of less than 2.3 Å, while the closest hydrogen to CO is from Leu 482 with a CO···H distance of more than 2.5 Å (Fig. 3b).

The structural features of the [NiFe] active site described above are in good agreement with recent density functional theory (DFT) calculations, performed on a large model including the second coordination sphere of the surrounding amino-acid residues using the ORCA program suite (Table 1)[28]. Geometries and infrared frequencies of the Ni-R state were obtained, with a hydride at the bridging position and a protonated terminal cysteine (Cys 546)[28]. The present experimental and theoretical results provide the structural basis for understanding the details of the heterolytic hydrogen splitting in the [NiFe] hydrogenase and deliver further insight into the electron and proton transfer. This

information is indispensable not only for understanding nature's ingenious enzymes but also for designing biomimetic or bioinspired model systems and devices for a future (bio)hydrogen technology[1].

1. Lubitz, W., Ogata, H., Rüdiger, O. & Reijerse, E. Hydrogenases. *Chem. Rev.* **114,** 4081–4148 (2014).
2. Volbeda, A. *et al.* Crystal structure of the nickel-iron hydrogenase from *Desulfovibrio gigas. Nature* **373,** 580–587 (1995).
3. Higuchi, Y., Yagi, T. & Yasuoka, N. Unusual ligand structure in Ni–Fe active center and an additional Mg site in hydrogenase revealed by high resolution X-ray structure analysis. *Structure* **5,** 1671–1680 (1997).
4. Ogata, H., Lubitz, W. & Higuchi, Y. [NiFe] hydrogenases: structural and spectroscopic studies of the reaction mechanism. *Dalton Trans.* 7577–7587 (2009).
5. Pandelia, M. E., Ogata, H. & Lubitz, W. Intermediates in the catalytic cycle of [NiFe] hydrogenase: functional spectroscopy of the active site. *ChemPhysChem* **11,** 1127–1140 (2010).
6. Armstrong, F. A. *et al.* Dynamic electrochemical investigations of hydrogen oxidation and production by enzymes and implications for future technology. *Chem. Soc. Rev.* **38,** 36–51 (2009).
7. Fontecilla-Camps, J. C., Volbeda, A., Cavazza, C. & Nicolet, Y. Structure/function of [NiFe]- and [FeFe]-hydrogenases. *Chem. Rev.* **107,** 4273–4303 (2007).
8. Higuchi, Y., Ogata, H., Miki, K., Yasuoka, N. & Yagi, T. Removal of the bridging ligand atom at the Ni–Fe active site of [NiFe] hydrogenase upon reduction with H₂, as revealed by X-ray structure analysis at 1.4 Å resolution. *Structure* **7,** 549–556 (1999).
9. Brecht, M., van Gastel, M., Buhrke, T., Friedrich, B. & Lubitz, W. Direct detection of a hydrogen ligand in the [NiFe] center of the regulatory H₂-sensing hydrogenase from *Ralstonia eutropha* in its reduced state by HYSCORE and ENDOR spectroscopy. *J. Am. Chem. Soc.* **125,** 13075–13083 (2003).
10. Fan, C. *et al.* Detection and characterization of exchangeable protons bound to the hydrogen-activation nickel site of *Desulfovibrio gigas* hydrogenase: a ¹H and ²H Q-band ENDOR study. *J. Am. Chem. Soc.* **113,** 20–24 (1991).
11. Whitehead, J. P., Gurbiel, R. J., Bagyinka, C., Hoffman, B. M. & Maroney, M. J. The hydrogen binding site in hydrogenase: 35-GHz ENDOR and XAS studies of the Ni-C active form and the Ni-L photoproduct. *J. Am. Chem. Soc.* **115,** 5629–5635 (1993).
12. Stein, M. & Lubitz, W. The electronic structure of the catalytic intermediate Ni-C in [NiFe] and [NiFeSe] hydrogenases. *Phys. Chem. Chem. Phys.* **3,** 5115–5120 (2001).
13. De Lacey, A. L., Fernández, V. M., Rousset, M. & Cammack, R. Activation and inactivation of hydrogenase function and the catalytic cycle: spectroelectrochemical studies. *Chem. Rev.* **107,** 4304–4330 (2007).
14. Afonine, P. V., Lunin, V. Y., Muzet, N. & Urzhumtsev, A. On the possibility of the observation of valence electron density for individual bonds in proteins in conventional difference maps. *Acta Crystallogr. D* **60,** 260–274 (2004).
15. Spatzal, T. *et al.* Evidence for interstitial carbon in nitrogenase FeMo cofactor. *Science* **334,** 940 (2011).
16. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
17. Gruene, J., Hahn, H. W., Luebben, A. V., Meilleur, F. & Sheldrick, G. M. Refinement of macromolecular structures against neutron data with *SHELXL2013. J. Appl. Cryst.* **47,** 462–466 (2014).
18. Lindahl, P. A. Metal-metal bonds in biology. *J. Inorg. Biochem.* **106,** 172–178 (2012).
19. Barton, B. E. & Rauchfuss, T. B. Hydride-containing models for the active site of the nickel-iron hydrogenases. *J. Am. Chem. Soc.* **132,** 14877–14885 (2010).
20. Ogo, S. *et al.* A functional [NiFe] hydrogenase mimic that catalyzes electron and hydride transfer from H₂. *Science* **339,** 682–684 (2013).
21. Weber, K. *et al.* A functional [NiFe]-hydrogenase model compound that undergoes biologically relevant reversible thiolate protonation. *J. Am. Chem. Soc.* **134,** 20745–20755 (2012).
22. Dementin, S. *et al.* A glutamate is the essential proton transfer gate during the catalytic cycle of the [NiFe] hydrogenase. *J. Biol. Chem.* **279,** 10508–10513 (2004).
23. Cleland, W. W. & Kreevoy, M. M. Low-barrier hydrogen bonds and enzymatic catalysis. *Science* **264,** 1887–1890 (1994).
24. Teixeira, V. H., Soares, C. M. & Baptista, A. M. Proton pathways in a [NiFe]-hydrogenase: a theoretical study. *Proteins* **70,** 1010–1022 (2008).
25. Galván, I. F., Volbeda, A., Fontecilla-Camps, J. C. & Field, M. J. A QM/MM study of proton transport pathways in a [NiFe] hydrogenase. *Proteins* **73,** 195–203 (2008).
26. Shomura, Y., Yoon, K. S., Nishihara, H. & Higuchi, Y. Structural basis for a [4Fe-3S] cluster in the oxygen-tolerant membrane-bound [NiFe]-hydrogenase. *Nature* **479,** 253–256 (2011).
27. Ogata, H. *et al.* Activation process of [NiFe] hydrogenase elucidated by high-resolution X-ray analyses: conversion of the ready to the unready state. *Structure* **13,** 1635–1642 (2005).
28. Krämer, T., Kampa, M., Lubitz, W., van Gastel, M. & Neese, F. Theoretical spectroscopy of the Niᴵᴵ intermediate states in the catalytic cycle and the activation of [NiFe] hydrogenases. *ChemBioChem* **14,** 1898–1905 (2013).
29. Pandelia, M. E., Infossi, P., Stein, M., Giudici-Orticoni, M. T. & Lubitz, W. Spectroscopic characterization of the key catalytic intermediate Ni-C in the O₂-tolerant [NiFe] hydrogenase I from *Aquifex aeolicus*: evidence for a weakly bound hydride. *Chem. Commun.* **48,** 823–825 (2012).
30. Rippers, Y., Horch, M., Hildebrandt, P., Zebger, I. & Mroginski, M. A. Revealing the absolute configuration of the CO and CN⁻ ligands at the active site of a [NiFe] hydrogenase. *ChemPhysChem* **13,** 3852–3856 (2012).

Author Contributions K.N. performed the sample preparation. K.N. and H.O. collected the X-ray diffraction data and performed the data analysis. H.O. measured the FTIR spectra. H.O. and W.L. designed and coordinated the project and wrote the manuscript.

Author Information Crystallographic data and coordinates have been deposited in the Protein Data Bank under accession numbers 4U9H (Ni-R_UH) and 4U9I (Ni-R_H), respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.O. (hideaki.ogata@cec.mpg.de) or W.L. (wolfgang.lubitz@cec.mpg.de).

METHODS

No statistical methods were used to predetermine sample size.

Sample preparation. The samples were prepared according to the methods reported previously[27]. All steps, including cell culture, purification of the enzyme and crystallization, were performed under strict anoxic conditions. For the crystallization, the vapour-diffusion sitting-drop method was applied. The buffer contained 50% 2-methyl-2,4-pentanediol in 25 mM Tris-HCl pH 7.4 and the protein drop contained 33% 2-methyl-2,4-pentanediol with 0.27 mM protein. A glove box (Coy Laboratory) was maintained under anoxic conditions with H_2 (1.5 ± 0.5%) and N_2 (98.5 ± 0.5%).

FTIR spectroscopy. The oxidation state of the enzyme was confirmed by FTIR spectroscopy. FTIR spectra were recorded on a Bruker IFS66v/S FTIR spectrometer with a 2 cm^{-1} spectral resolution at 293 K. Data collection and baseline correction used OPUS software (Bruker). The FTIR spectra are shown in Extended Data Fig. 2.

Diffraction data collection. The crystal (typical of 2.0 mm × 0.2 mm × 0.2 mm) was mounted on a 0.4 mm cryo-loop (Hampton Research) without a reservoir solution to decrease background noise. During the diffraction experiments, the rather large crystal was translated slightly along the rotation-axis direction after each rotation angle of 30°. At the BESSYII BL14.2 beamline (Helmholtz-Zentrum Berlin), two complete data sets were collected. Eight diffraction sets from a single crystal (termed 'Ni-R_{UH}') were merged and scaled up to 0.89 Å resolution by XDS and XSCALE[31]. For 'Ni-R_H', three diffraction sets were merged and processed up to 1.06 Å resolution. The resolution limit was chosen by the cut-off criteria of $<I/\sigma> > 1.0$ and $CC_{1/2} > 0.5$ in the outer-shell[32]. The intensities of each data set were scaled to the highest-resolution diffraction data set. The reflection data were converted to HKL4 format by XDSCONV[31]. Data collection statistics are summarized in Extended Data Table 1.

Structure refinement. The molecular replacement method was applied by using the initial model (PDB accession number 1WUI)[27]. Model building and refinement were performed by COOT and PHENIX, respectively[16,33]. After the molecular replacement (by PHASER at 3.5 Å resolution), the following refinement strategy was applied. First, rigid-body refinement was performed at 1.5 Å resolution with the hydrogen atoms at the amino-acid residues as the riding model. The isotropic B-factor was set to 15 Å^2 and the water molecules were added at 1.5 Å resolution. Then, the resolution range was extended to 1.1 Å resolution by steps of 0.2 Å after each three refinement cycles. At 1.3 Å resolution the isotropic atomic displacement parameters refinement for all atoms was applied. Then, the anisotropic atomic displacement parameter refinement of all atoms except for the hydrogens was applied at 1.1 Å resolution. In the next step at 1.0 Å resolution, the occupancies of the water molecules were refined. In the last step, the multiconformations of the amino-acid residues were checked and modified in the model manually. During the refinement the Friedel-pair reflections were treated as different, and the anomalous effects were refined in case of the 12 iron atoms, 1 nickel atom and 45 sulphur atoms. Finally, the refinement was performed up to the maximum resolution. During the refinement up to 1.0 Å resolution, the N-Gaussian (a dynamic adjustment of the number of Gaussian terms) model was applied to calculate the scattering factors. At the highest resolution, the Wk1995 model (the 4-Gaussians and a constant) were applied for the scattering factors. For the metal ions, the scattering factors for Mg^{2+}, Fe^{2+} and Ni^{2+} were used. The electron density maps were calculated with the ρA-weighted map ($2mF_o - DF_c$ filled with the missing reflections, and $mF_o - DF_c$ difference

map). The $F_o - F_c$ omit maps in the figures were calculated using the models without the hydrogen atoms. Structure refinement statistics are summarized in Extended Data Table 1.

Hydrogen modelling. The hydrogen atoms of the amino-acid residues were modelled as a riding model which was generated automatically with the program PHENIX. After each refinement step they were checked manually, for example the orientations of the methyl groups, the hydroxyl groups (Tyr, Ser and Thr) and the protonation states of the His residues. Moreover, the hydrogen atoms of the water molecules were also modelled and were checked manually when they were visible. In the $F_o - F_c$ omit map, the weak residual electron densities in the [NiFe] active site and at the cysteine residue (Cys 546) were assigned to a hydride and a protonated cysteine, respectively. The position of the hydride was modelled based on the $F_o - F_c$ omit map. The refinement of these atoms was performed with distance restraints. The electron density analysis used MAPMAN[34]. The hydrogen–carbon (or nitrogen) distances of the riding model for neutron diffraction in the Extended Data Fig. 3b were taken from ref. 17.

Validation of the refined model. Stereochemical properties were checked by a Ramachandran plot from PHENIX. Estimated coordinate errors (except for the hydrogens) were calculated by least-square full-matrix inversion using the program SHELXL[35]. Figures were generated by PyMol (http://www.pymol.org/).
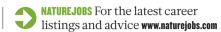
Sequence alignment. Sequence alignment of the [NiFe]/[NiFeSe] hydrogenases was by CLUSTAL Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). The sequences of the [NiFe]/[NiFeSe] hydrogenases were taken from the PDB. The figure of the sequence alignment was generated by ESPript3.0 (http://esprit.ibcp.fr).
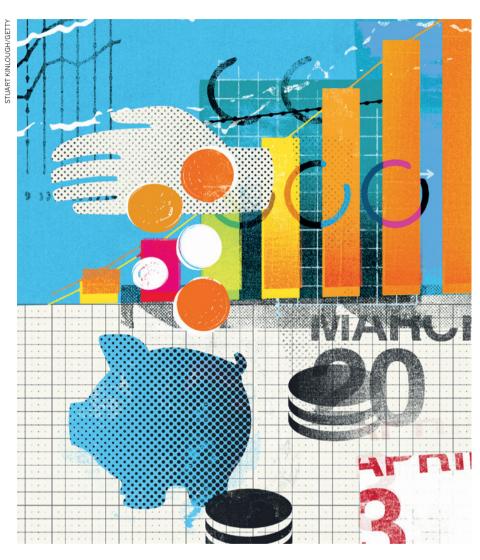
31. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
32. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).
33. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
34. Kleywegt, G. J. & Jones, T. A. xdlMAPMAN and xdlDATAMAN - Programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Crystallogr. D* **52,** 826–828 (1996).
35. Sheldrick, G. M. A short history of *SHELX. Acta Crystallogr. A* **64,** 112–122 (2008).
36. Breitenfeld, J., Scopelliti, R. & Hu, X. Synthesis, reactivity, and catalytic application of a nickel pincer hydride complex. *Organometallics* **31,** 2128–2136 (2012).
37. Gehring, H. *et al.* Hydride reactivity of Ni^{II}-X-Ni^{II} entities: mixed-valent hydrido complexes and reversible metal reduction. *Chem. Eur. J.* **19,** 1629–1636 (2013).
38. Zhang, X., Huang, D., Chen, Y.-S. & Holm, R. H. Synthesis of binucleating macrocycles and their nickel(II) hydroxo- and cyano-bridged complexes with divalent ions: anatomical variation of ligand features. *Inorg. Chem.* **51,** 11017–11029 (2012).
39. Huang, D. & Holm, R. H. Reactions of the thermal Ni^{II}-OH group in substitution and electrophilic reactions with carbon dioxide and other substrates: structural definition of binding modes in an intramolecular Ni^{II}···Fe^{II} bridged site. *J. Am. Chem. Soc.* **132,** 4693–4701 (2010).
40. Lee, C. H., Lutterman, D. A. & Nocera, D. G. Photoactivation of metal-halogen bonds in a Ni(II) NHC complex. *Dalton Trans.* **42,** 2355–2357 (2013).
41. Bau, R. & Drabnis, M. Structures of transition metal hydrides determined by neutron diffraction. *Inorg. Chim. Acta* **259,** 27–50 (1997).
42. Vicic, D. A., Anderson, T. J., Cowan, J. A. & Schultz, A. J. Synthesis, structure, and reactivity of a dinuclear metal complex with linear M-H-M bonding. *J. Am. Chem. Soc.* **126,** 8132–8133 (2004).
43. Chotard, J.-N., Filinchuk, Y., Revaz, B. & Yvon, K. Isolated $[Ni_2H_7]^{7-}$ and $[Ni_4H_{12}]^{12-}$ ions in $La_2MgNi_2H_8$. *Angew. Chem. Int. Ed.* **45,** 7770–7773 (2006).

CAREERS

STUART KINLOUGH/GETTY

COLUMN
Respect the report

Foundations have reporting requirements that must be followed, notes grant-director **Ingrid Eisenstadter**.

When a grant applicant receives an early-career award from a private foundation, she or he enters into a crucial period of communication with the sponsor that will last at least a year, often longer. Foundations — like most granting organizations — have exacting reporting requirements. Grant recipients will have to submit at least one report on the progress of their work, depending on the length of the sponsored research and whether the project changes direction at all.

The recipient of a one-year grant, for example, might need to submit a progress report after six months and a final report once the project is completed. The interim report might be little more than a progress update, whereas the final report will be comprehensive and detailed. Multiple-year grants could call for a single report on completion of the research or successive annual ones. Every foundation has different requirements. And if the research protocol changes or the work requires more time, the researcher should seek approval from the granter as soon as the situation arises.

Foundation-grant recipients need to familiarize themselves with relevant reporting requirements, including formats, deadlines and expected content. These might be spelt out in the covering letter that announces the award, or could be found on the foundation's website. If neither offers answers, grant recipients should consult their contact at the foundation, if they have one, or the person who signed the letter. Failure to follow the foundation's guidelines (or failure to meet a deadline) could jeopardize the recipient's standing with the funder for future proposals (see 'Golden rules for reporting').

PLAN YOUR PATH

As a grant recipient, you should view the granter's reporting requirements as a map for how to write your reports. A foundation may require researchers to use a prepared template or it may allow you to use your own format. If you are not using a template, take care not to fall into the 'copy-and-paste' habit. From my experience as director of grants for a US family foundation in the sciences, I find that our grant recipients sometimes insert great swaths of their original proposals into their final reports. I suspect that recipients do this to bulk up the report and make it look more thorough.

It does not work. It is clutter, and usually heralds an abbreviated discussion of the actual results and accomplishments. Instead, briefly summarize your work and highlight your victories at the outset, and then explain them in detail in the following pages.

A sensible way to proceed is to address each subsection as it appeared in your proposal. If your proposal listed steps such as 'select a 40-patient cohort, collect weekly blood samples, and sequence viruses', repeat these headings in your report. Under each heading, describe your progress.

Your final report should end with your ▶

▶ conclusions and their ramifications. If you hope to apply for future support, you might also include 'next steps'.

CHANGE HAPPENS

Foundations understand that delays can happen: a shortage of supplies, delayed permits for travel or fieldwork, or a team member who has unexpectedly relocated. When this occurs, you should promptly notify your granter and ask for a no-cost extension. This is a fairly common request that requires no additional outlay from the foundation, which will probably not object. In my 25 years at the foundation, we have never denied such a request.

If you are forced to modify your original research protocol, you must notify the foundation. Such a request is unlikely to present problems; we have never denied a request for this type of change. We expect researchers to know better than we do how they should react to an unexpected lab-test result, for instance.

But a major change could cause trouble. If the foundation determines that you are no longer carrying out the work that it wants to support, it could suspend or rescind your grant. I recall only two times when we did not receive prior notification of a major change in research plans. In one case, an animal-migration researcher did not carry out the task that was most important to us — an examination of the harm caused by underwater blasting in oil exploration. In the other, a mouse trial was scaled back to such an extent that it had no hope of progressing to a human trial in the foreseeable future. We viewed the researchers' failure to notify us of such substantial changes as a reflection of lax internal administrative management. We have not funded either organization since then.

You could run the same risk. So before you request a major change in protocol or extra funding, ask colleagues if they know how the foundation operates in those circumstances, or seek advice from your mentor or development office. In rare circumstances, your institution could make up a funding shortage.

ATTENTION TO BUDGET

Development-office involvement in grant proposals and reports varies substantially between institutions. In some, development staff have little to do with grant applications or reports; in others, they work with research teams to write their grant proposals and reports. If you find yourself depending on these other teams, review their work carefully. Usually, people who do this work are not graduate-level scientists and can introduce mistakes or omit crucial information.

It is important to cross-check reports of your expenditures against the budget you originally submitted and to explain any minor reallocations — which are commonplace — for which you did not seek advance approval. If you reallocate a relatively small amount — US$500, or even $1,000, in a $25,000 budget — you

probably do not need advance approval. If you are unsure, check with your granter first.

Some years ago, we received final reports for two virus studies that we had funded simultaneously at the same organization. The researchers there had not written their own grant proposals or reports; the development office had. Nor had research teams created their budgets; the finance office had. Yet no one had considered the original budgets in the final expense reports: not a single budget line in either recipient's report agreed on cost or designation. As we read the final reports, we could not determine how our funds had been spent. We have not funded that organization since.

If you have leftover funds on completion of your project, seek your granter's permission to hold onto those funds, and explain how you will spend them. Do not assume you can keep the money. Not surprisingly, it is unusual for us to see final reports that indicate surplus funds. Indeed, we sometimes receive final budget reports that agree line for line with the original budget, down to the last dollar. Such tight adherence to budget would require the intervention of the tooth fairy, but it is fairly common. We assume that it resembles reality closely enough.

Writing a final report often coincides with preparations to submit the research for publication. Occasionally, we learn that researchers we have supported have had the first phase of their work published by the time we received their final reports. If this is the case for you, include this information in your final report

and, as a courtesy, send links or copies of your future publications to your funder as they appear. Publications in peer-reviewed journals are important to us — we see them as a confirmation of the significance of the work.

RISKY BUSINESS

From time to time, we receive final reports that ask for renewed funding at the same time. Although researchers may request permission to submit a subsequent grant application with their report, occasionally, they include a new proposal. Jumping the gun in this way carries risk, yet it is understandable. Although some private foundations meet monthly, others meet only semi-annually or annually; waiting for the next formal submission date could substantially delay your follow-up application.

Should you take this initiative? It is a tough call. An unsolicited proposal submission might be considered too forward or the foundation might have a policy not to award another grant until final reports from the previous round have been circulated and reviewed (a process that could take months). If you are in any doubt, contact your sponsoring foundation.

Do not drown your granter with enquiries — always search the foundation's website first. But the occasional request for clarification will reduce your risk and save time for applicant and granter alike. ■

Ingrid Eisenstadter *is director of grants for the Eppley Foundation for Research in New York.*

GOLDEN RULES FOR REPORTING
What grant recipients should be doing now

Private foundations might make recurring grants over the years to researchers whose work shows substantial promise. To keep that door open, grant recipients should keep caveats in mind as they create their progress reports.

REMEMBER

● Speak English. Avoid highly technical language. Private foundations in particular might have non-scientists on their boards and you want them to understand what you have accomplished.

● Circulate to colleagues. Ask co-workers to read reports for content and errors, and get feedback on how well you have captured the importance of your work and whether you have clearly highlighted the significance of your results.

● Future ramifications. No one wants to fund work that will simply sit unread in a library. Explain how your results will drive your future research or that of others.

● Format thoughtfully. Headings and subheadings provide clarity and guidance. Include a table of contents if the report is long.

AVOID

● Lateness. Put due dates for interim and final reports in your calendar as soon as the grant is awarded. Add a reminder a month ahead of time and go to the foundation's website to make sure that it has not changed its reporting requirements.

● Obscuring the point. Do not start your interim and final reports with a lengthy rehashing of background that was already in your grant proposal. Keep the rest of the report succinct and on point to avoid burying your actual achievements when you should be highlighting them.

● Omitting conclusions. State clearly what you learned from your project. If your results were ambiguous or not what you anticipated, that is not a crime. Say so, and describe what you might do in the future to get clarification. **I.E.**

TRANSFERENCE

The shock of the new.

BY JENNIFER CAMPBELL-HICKS

I come to consciousness gently after transference, like rising from calm, deep water. I open my eyes. The recovery room is a washed-out blue with yellow undertones, as if I'm looking through a filter. A strong scent of lemon cleaner assaults my nose. After-effects of the procedure, I'm sure. I'm curled in a pillowed bed.

A middle-aged woman leans over me. She's wearing a lab coat, and her long blonde hair is tied back. Her scent is bitter and antiseptic. I decide she must be a doctor.

"Welcome back, Ms Bragin," she says. "How do you feel?"

I lift myself up and stretch, long and languid, from head to toes. How do I feel? Strong. Like I could run forever. It's much better than before I died, when my old body had wasted into weakness.

"Wonderful," I say.

The word doesn't come out as a word. It's a bark.

Startled, I jerk to my feet.

I look down. I have four paws, claws and golden fur. I look back. I have a tail that I have tucked between my legs.

Oh, no. No, no, no.

My will specifically stated that after my death, I wanted transference into the body of a god. I didn't specify which one. Greek, Norse, Egyptian, Hindu, whatever was available. Gods are ridiculously expensive. They take years to grow in the lab. But I had been filthy rich. What else would I do with the money, if not buy comfort in my next life?

Beautiful, pampered and adored. That's what I wanted.

Instead of god, I got dog.

"There's been a mistake!" I try to say, but my complaint emerges as a pathetic whine.

Who did this to me? Did my family misread my will? No, they would never have made such an error. Did my lawyer type it wrong? That's more likely. He always disliked me, and any typo of his might have been intentional. I bet he's laughing now.

"It's all right, Ms Bragin," the doctor says. "It can take a few minutes to adjust to your new condition,

but I promise you'll feel better soon." She scratches behind my ears, and oh, it's like heaven. My eyes drift shut. Despite myself, I lean into her hand. "Our canine clients have the highest satisfaction rate with their new life. Ninety-five per cent. Much better than cats."

ILLUSTRATION BY JACEY

She strokes my fur, gives me a pat and walks across the room to her computer. I shake myself from her spell. I can't let myself be seduced. I need to tell her this is a mistake, but how? I've already established that I can't speak, and I can't write, either. No thumbs. I could try typing a message on the doctor's computer, if I could reach the keyboard and if my paws could tap one key at a time.

While I consider the feasibility of this option, the doctor comes back. "Good news, Ms Bragin. Your niece is here to pick you up. Cindy Mcree."

I perk up. Cindy is here? Surely she can fix this. She must have seen my will and understood what I wanted, no matter what my lawyer wrote. She'll tell the doctors, and next time I wake up, I'll be Athena or Devi.

I wait. Soon, Cindy walks in, dressed impeccably, a queen of fashion and good taste. She smells like her favourite perfume and smiles when she sees me. I wag my tail. Tell the doctor, I think. Tell her they got it wrong.

She cups my muzzle in both hands.

"Aunt Penny," she says. "Look at you. You're so beautiful, just like you always wanted. I know how much you loved golden retrievers. What lovely fur you have."

She strokes my back, from neck to tail. I shiver with delight, and my tail wags harder. I'm beautiful?

Wait. I don't want this.

Or do I?

My treacherous tongue is lolling. I pull it into my mouth and snap shut my muzzle.

Tell the doctor, I think, but my conviction is wavering.

"Let's get you out of here," Cindy says. "I'll take you to the park so you can test your new legs and chase some squirrels. When we get home, I'll have the cook make you a juicy steak."

My legs twitch, and my mouth waters. I have to admit, that sounds glorious. Maybe this isn't so bad. I could do worse. Like the doctor said, I could have ended up a cat. The park has plenty of squirrels in need of chasing, and I could breathe in fresh air instead of this sterile clinic smell. I could sniff the trees and add my own scent to introduce myself to the neighbourhood's canines.

Beautiful, pampered and adored. That's what I wanted. Who's more pampered and adored than a rich and beautiful dog?

I nuzzle at Cindy to show my agreement.

"Come on, Aunt Penny," she says.

I leap from the bed to follow her, my claws clicking on the linoleum floor. I can't wait to dig them into cool grass and dirt while I chase a squirrel up a tree. ∎

Jennifer Campbell-Hicks *is a writer, journalist, wife, mother and lifelong fan of science fiction and fantasy. Her fiction has appeared in* Daily Science Fiction, Flash Fiction Online *and* Intergalactic Medicine Show. *She blogs at jennifercampbellhicks. blogspot.com.*